Reference and quantification in the cognitive view of language.

Han Reichgelt

PhD
University of Edinburgh
1985

The research reported in this thesis is my own and the thesis has been composed by myself.

Han Reichgelt.

Acknowledgements.

I would like to thank my supervisors Prof. Dr Barry Richards and Dr Mark Steedman for their advice and many helpful comments on earlier drafts. I also would like to thank Dr Kit Fine for his comments on chapter 5. A lot of the ideas in this thesis resulted from discussions with Nigel Shadbolt and his help is gratefully acknowledged. Over the years I have given a number of informal talks to various workshops in the then School of Epistemics. Various members of the School have helped me by giving their critical comments. Finally, I would like to thank my examiners Dr Phil Johnson-Laird and Dr Henry Thompson for pointing out some problems with an earlier version of this thesis. Needless to say, any remaining mistakes are entirely my responsibility.

I would also like to thank the Niels Stensen Stichting in Amsterdam and the University of Edinburgh for giving financial support for the research reported here. I also like to thank the Edinburgh University Speech Input Project for letting me use their work-processing and printing facilities. At least the thesis looks good. Finally, I would like to thank Ms Nicola Cowie for her moral support and for urging me to submit my thesis as soon as possible.

Abstract.

In this thesis I take the cognitive view of language according to which language has to be studied in terms of the processes occurring in the minds of speakers and hearers when they are producing or understanding discourses. According to this view, the explanation of linguistic phenomena also has to be in terms of the mental representations of discourse used in language production and constructed in comprehension. Another consequence is that a satisfactory theory has to provide a theory both of the architecture of the human language processing mechanism and of the detailed representations it uses or constructs in discourse.

I provide a theory of the architecture of the human language processor. The central notion of the notion of knowledge activation. The model postulates three different components which represent different degrees of knowledge activation. Another important idea is the idea of embedded models in each of these components, which are used to represent the beliefs and knowledge the language processor ascribes to his fellow discourse participants.

The detailed representations which are used or constructed in discourse are postulated to be frame based in the sense in which this notion has been developed in Artificial Intelligence. I define an expressively somewhat impover- ished formal knowledge representation language for which I define a direct model-theoretic semantics using the theory of arbitrary objects developed by Fine. The fact that my knowledge representation formalism has a formal semantics sets it apart from the majority of alternative frame based representation languages.

Finally, I sketch analyses of a number of referential and quantified expressions in English. I use both the proposals made for the overall architec- ture of the human language processing mechanism and for the detailed representa- tions it uses or constructs, thereby illustrating their potential usefulness in the explanation of the behaviour of certain expressions in English.

Chapter 1:- Introduction

A lot of work in the area of natural language is based on the belief that the mental processes which go on when one is learning or employing certain concepts ought not to play a role in the analysis of these concepts. This is evident in the formal semantics approach to language, as started by Carnap (1947) and further developed by Montague. It also underlies the transformational approach to syntax. Notwithstanding Chomsky's insistence that

> linguistic theory is mentalistic, since it is concerned with discovering a mental reality underlying actual behavior.
>
> Chomsky (1965,4)

transformational linguists never ask psychological questions (McCawley, 1982). In other words, the Chomskian position also seems to be that one can study the structure of language independently of the cognitive processes going on in speakers and hearers who use language.

However, with the emergence of cognitive science the opposite view has become more popular. Many, especially in Artificial Intelligence, have argued that the study of the semantics of natural language has to take processing factors into account (Cf. Winograd, 1976). Lakoff (1982) reviews some of the examples in the literature that indicate that in both syntax and semantics one has to take into account the processing of sentences as they are uttered. He concludes:-

> After a generation of research in which it was implicitly assumed that language could be described in its own terms, it has become more interesting to ask how much of the structure of language is determined by the fact that people have bodies with perceptual mechanisms and memory and processing capabilities and limitations, by the fact that people have to try to make sense out of the world using limited resources, and by the fact that people live in social groups and have to try to communicate with each other. It seems to me that a great deal of the structure of language is determined by such factors.
>
> Lakoff (1982,155)

This thesis is an attempt at providing a theory in which some of the structure of language can be explained in terms of the factors Lakoff mentions. I will provide a theory of the human language processing mechanism, or processor for short, and attempt to show that some of the structure of language can be more lucidly described in this theory.

By adopting the cognitive view of language, one does not necessarily reject work based on the formal view of language. The issue at stake is not a better solution to the problems as they have been posed, but rather a redefinition of the nature and the structure of the relevant problems, as Winograd (1976,282) points out. The claim is that the redefinition of the problems from the cognitive point of view will enable us to highlight and clarify some aspects of the problems which remain obscure under the old definitions.

The outline of this thesis is as follows. In chapter 2, I briefly discuss the cognitive view of language and its implications for theories of the human language processor. In particular, I argue that a complete theory has to provide both an overview of the overall structure of the processor (its architecture if you like), and of the detailed structures it creates in language understanding, or uses in language production.

In chapter 3, I review some of the theories which have been proposed in the literature. I discuss two proposals for the overall architecture of the processor, and a number of proposals concerning the detailed representations used in discourse. On the basis of my discussions of the latter I conclude that the representations used in discourse are structurally identical to the representations used to store long-term knowledge. The question of the structure of discourse representations thus is identical to the question of knowledge representation.

In chapter 4, I introduce my own model of the overall architecture of the language processor. The model I propose is very similar to that of Sanford and Garrod, one of the proposals discussed in chapter 3. Two notions will play a cental role:- the notion of knowledge activation and the notion of embedded models which can be used to represent the beliefs one language user ascribes to another.

In chapter 5 I provide a knowledge representation scheme which describes the detailed structures used in the language processor. The knowledge representation system is very similar to but admittedly less expressive than a number of existing knowledge representation schemes. The major innovation lies in the fact that I provide a direct denotational semantics for parts of the system. One of the most important aspects of the proposed knowledge representation scheme is the fact that knowledge is stored in relatively large chunks.

In chapter 6 I first discuss the notion of reference and I then use the model developed in chapter 4. and 5. to discuss the behaviour of some expressions in English. The expressions which are analysed are indefinite NPs, definite descriptions, pronouns, universally quantified NPs and quantified NPs including the so-called non-standard quantifiers *most* and *few*. Chapter 6. is intended as a practical argument for

the viability of the cognitive view of language and procedural semantics in general and the more specific proposals made in this thesis in particular. The analyses which I propose rely heavily on the proposals made in chapters 4. and 5. and to the extent that they throw some light on some of the open problems in the semantic analysis of the expressions discussed, to that extent can the cognitive view of language be said to have proved its usefulness.

Chapter 7. is a critical discussion of the achievements and shortcomings of this thesis. In it I also discuss some ways in which the present thesis could be expanded.

In this thesis I will often have to use the term 'speaker' and ´hearer'. I will assume that the speaker is female while the hearer is male. The personal pronoun 'she' will thus always denote the speaker while the personal pronoun 'he' stands for the hearer.

Chapter 2:- Philosophical background

1. Introduction

In this chapter I will first briefly discuss the philosophical view of language underlying the present approach to language. I will then defend it against a number of objections which might be raised against it. Finally, I will discuss some of the consequences which this philosophical view has for theories of language and the human language processor.

2. Procedural semantics

What I have called the cognitive view of language, is also known under the name of procedural semantics. Maybe the best summary of this view of language can be found in Winograd (1976, 262-63). Winograd gives a list of six assumptions underlying the procedural view of language, which I will call the axioms of procedural semantics. They are:-

> Axiom 1. The primary focus in the study of language should be on the mechanisms underlying language production and comprehension.
>
> Axiom 2. The essential properties of language reflect the cognitive structure of the human language user.
>
> Axiom 3. Language use takes place within a structure of ongoing thought processes.
>
> Axiom 4. Each utterance is constructed to serve a combination of communicative goals.
>
> Axiom 5. The most appropriate formalisms for building theories of language are those that deal explicitly with the structure of knowledge and the processes using it.

Axiom 6. There is a set of structures modified by speakers and hearer in the course of communication, and the theory of language use has to deal with a succession of structures and the nature of the changes

Thus, according to the cognitive view of language, language cannot be studied independently of the people using it. We have to study language as part of the ongoing thought processes of speaker and hearer. We are therefore primarily interested in the mechanisms underlying language production and comprehension. This involves the study of the overall architecture of the human language processor. One may for example want to distinguish between various parts of human memory which can be shown to play different roles in language understanding. However, we will also have to study the structure of the representations which are built up during language use, and the way in which they are constructed in response to certain expressions in English. One of the questions is for example the difference between an utterance of an indefinite NP as opposed to one of a proper name on the representation the hearer is constructing, or the related question under what circumstances the speaker will use an indefinite rather than a proper name. The fifth axiom claims that the most appropriate formalisms here are those which can also be used in the more general area of knowledge representation. We will return to this point in the next chapter.

There are a number of problems associated with this view of language. In the next two sections I will describe these problems, and sketch possible replies to them.

2.1. Intersubjectivity

A first problem for this cognitive approach to the study of language, is the

intersubjectivity of language. In general, we are reasonably successful at communicating whatever it is we want to communicate. We get our messages across and we understand what information others try to communicate to us. Since I cannot explain the success of language in terms of something external to the language user, the question arises how the intersubjectivity of language can be accounted for. What guarantee is there that the structures one language user constructs or uses when understanding or producing a piece of discourse are in some sense similar to the structures of another language user.

The answer to this question is based on the work of the naturalistic epistemologists. They have put forward some arguments for the trustworthiness of our perceptions. Without wanting to go into this philosophical doctrine in detail, I just summarise the argument very briefly:- the human species is very successful in evolutionary terms and there is therefore sufficient ground for assuming the trustworthiness of its cognitive system. After all, the evolutionary success of the human species indicates its success in coping with its environment and therefore one can suppose that there is a very close connection between the way the world is and the way the human species perceives it. This evolutionary explanation of the structure of the human cognitive system has the consequence that the structure is (at least partly) genetically determined. If the human cognitive system has been determined by its evolutionary development, then its structure has to be 'wired in' in each individual member of the species. Indeed, the earlier stages in the development of the human cognitive system as studied by Piaget and Bower and their associates, are very much the same for all members of the species. The structure of the human cognitive systems thus can be expected to be very similar among human beings, and success in communication can be explained on the basis of this fact. We all have the

same basic cognitive make-up and we deal with incoming information in very similar ways. What actual information there is in the individual cognitive systems depends on the experience the individual has had, but the way experiences are processed is by and large identical for all individuals. I will call this hypothesis *the Thesis of Cognitive Similarity.*

The Thesis of Cognitive Similarity is important not only on this theoretical level. It is also relevant in our dealings with each other. When we explain the behaviour of our fellow human beings, we take what Dennett (1978) calls 'the intentional stance'. We predict the behaviour of another person

> by ascribing to the system *the possession of certain information* and supposing it to be *directed by certain goals,* and then by working out the most reasonable or appropriate action on the basis of those ascriptions and suppositions.
>
> Dennett (1979,6)

The problem with this idea is how one works out what the 'most reasonable and appropriate action' is. In the processor-centric view there is only one possible answer:- you decide on the most reasonable action by working out what you would do yourself, if you yourself held the beliefs and goals ascribed to the individual whose behaviour you want to explain.

Occasionally, the Thesis of Cognitive Similarity, and its collorary that everybody is rational, have to be given up in face of the facts. If someone consistently behaves in an unexpected (or irrational) way, one will give up the assumption that the person in question is a rational person. In Dennett's terms, we give up the intentional stance and take the design stance.

The Thesis of Cognitive Similarity plays an even more crucial role in language use. Without it, communication would be impossible. Searle (1969,16) writes:-

> When I take a noise or a mark on a piece of paper to be an instance of linguistic communication, as a message, one of the things I must assume is that the noise or mark was produced by a being or beings more or less like myself and with certain kinds of intentions.

So the Thesis of Cognitive Similarity is used on two different levels. First, we use it in our dealings with each other. Secondly, on a theoretical level an evolutionary argument can be advanced to show that this assumption is warranted and it can thus be used to explain the successfulness of communication in a purely cognitive approach to language use [1].

2.2. Language and the world

Another problem for the cognitive approach to language is the fact that people often use language to transfer information about the world. But if using language involves the construction of mental representations of some sort, then the problem arises how one can transfer information about the world. Again, the thesis of Cognitive Similarity can provide an answer to this objection.

Comprehending a piece of discourse often involves the activation of already existing knowledge and the transmission by the speaker of knowledge which is new to the hearer. The hearer is aware of the close relationship between most of his own knowledge and the world. Because of the Thesis of Cognitive Similarity, he will also assume this close relationship in the speaker. Therefore, if the hearer takes the speaker to be authoritative and cooperative (and to have the intention to talk about real-world objects), he will in general also assume that the new information he

receives from the speaker will bear the same close relationship to the world as his own knowledge. Conversely, the speaker assumes that if the hearer takes her seriously, the hearer will treat the information she is encoding in the utterance in the same way as she herself does. Therefore, she assumes that the hearer will in general take the information she is providing him with as applying to the real world as well.

3. Consequences for theories of the language processor

As said before, according to the cognitive view of language, a theory of language has to study primarily the mechanisms underlying language comprehension and language production. It follows that a complete theory of language consists of two parts. First, one will have to provide an overall theory of the structure of the human language processor i.e. a theory of its architecture, and theories developed in this area will have to be in accordance with psycholinguistic findings about language understanding and production. Secondly, a complete theory of the processor will also have to provide a theory about the detailed structures which it uses when generating a piece of discourse, or constructs when understanding a text. Clearly, theories in these areas also have to take into account the psycholinguistic findings which were relevant to theories about the architecture of the human language processor. However, since one is also making claims about the effects of certain linguistic expressions, or the conditions under which the speaker will use them, one will also have to take into account linguistic findings about the distribution of the expressions in question in discourse. There are thus two distinct types of criteria of adequacy on theories of the human language processing mechanism, linguistic and psycholinguistic ones. I will discuss the psycholinguistic criteria first.

3.1. Psychological criteria of adequacy

As a psychological theory, an adequate theory of the mental representation of discourse has to be compatible with at least four rather basic psychological facts.

3.1.1. The 'veil of perception'

The first psychological fact is the 'veil of perception' (Locke, 1690). Locke observed that human language processors do not have direct access to the world but only have their mental representations of the world. They can only get in contact with their environment through their cognitive systems. Even our perceptions are mediated through the structure of our cognitive systems. Modern psychology has proved Locke right:- it has shown that the view of our cognitive systems as directly mirroring the world around us has to be given up. An adequate theory of the human language processor should respect this fact.

3.1.2. The finiteness of our cognitive systems

A second basic psychological fact which a theory of the human language processor has to take into account is that our cognitive capacities are finite. Thus, the mental representation discourse processors construct in response to a discourse have to be finite and have to be constructable in a finite amount of time. A theory that postulates infinite structures as an integral part has to be rejected.

3.1.3. Left-to-right processing

A third psychological criterion which has to be met is a consequence of the claim that processing takes place on a left to right basis. The claim can be made in various forms but I will use it in its weakest form. The claim is that sometimes some part of

the mental representation of an utterance will have been constructed before the processor encounters the end of the utterance. Experimental evidence is presented by Crain (1980), Crain and Steedman (1983) and Altmann (forthcoming). Previous discourse and other information of a 'semantic' and/or 'pragmatic' kind can heavily influence the outcome of the parsing process. Crain and Steedman conclude:-

> The results suggest that there is no such thing as an intrinsically garden pathing sentence *structure*, but rather that for a given sentence, certain *contexts*, (possibly including the null context) will induce a garden path effect, while others will not.

A satisfactory theory about the human language processor thus should be compatible with the fact that the human language processor is able to do some semantic work before the entire utterance has been parsed.

3.1.4. The relevance of background knowledge

A final psychological fact which I take to be of central importance to any theory of the mental representation of discourse is the relevance of background knowledge. There is ample psychological evidence showing the relevance of background knowledge both in text comprehension and text recall. I do not want to give an extensive review of the psychological literature on text recall, or discuss the methodologies used by the various researchers in detail. I will rather concentrate on the main conclusions from their experiments. Moreover, I will restrict myself to those papers which deal with 'normal' use of language as opposed to those which deal with for example verbatim memory.

One of the first illustrations of the relevance of background information in text recall and text comprehension can be found in Bartlett (1932). He presented British subjects with red Indian folk stories. Obviously, these stories came from a

very different culture and hence were hard to understand. When Bartlett asked his subjects to reproduce the stories, they tried to "make sense out" of the story. They omitted details which they found incomprehensible or included certain sentences which made for example the causal connections between different episodes in the stories clear from their cultural perspective. People thus do not remember the stories only on the basis of the information in the actual text, but also use their background knowledge.

As further support of the relevance of background knowledge in text recall and comprehension, consider the following. If one uses two different referring expressions which people know to stand for the same object, because of their background knowledge, then there is a certain confusion in a recognition task as to which expression was used in the original sentence. Anderson and Bower (1973,248-52) for example presented subjects with a list of sentences which included sentences like (1):-

(1) The first president of the United States had bad health.

In a later recognition task, people were also presented with sentences in which the definite description was replaced by a co-referring proper name. So, people were presented with sentences like (2):-

(2) George Washington had bad health.

Anderson and Bower found that there was a confusion as to which term the original sentence contained and people often said that sentence (2) actually occurred in the list which was presented to them.

Sulin and Dooling (1974) found a similar interference from background knowledge in another recognition task. Two groups of subjects were presented with the same story, except that in one conditions the story contained an arbitrary name, whereas in the other it contained a well-known name. So, in one story they used the name 'Carol Harris' and in the other 'Helen Keller'. They found that people who were presented with the version of the story in which the name 'Helen Keller' was used, wrongly claimed that sentence (3) had occurred in the story, while the other group correctly claimed it had not.

(3) She was deaf, dumb and blind.

Both Anderson and Bower's and Sulin and Dooling's results indicate that even in recognition tasks the influence of background knowledge is immense. Background knowledge and information from the actual text combine to form a mental representation of the text. The afore-mentioned results indicate that background knowledge can actually override the information from the text itself.

A final illustration of the relevance of background knowledge for the comprehension of texts is provided by the experiments done by Bransford and Johnson (1972,1973) They showed that if one gives subjects some clues about what background knowledge is relevant for the comprehension of a particular text, either by giving cue words or titles, or by showing pictures, or by giving the story a title which was suggestive of the background against which the story was to be understood, both comprehension and memory for the text improved dramatically. Garrod and Sanford (1982) report similar results in a comprehension task:- reference to an object which would be expected to be present because of the title of a story is just as easy to understand independently of whether it has been introduced explicitly

before or not, whereas the same reference takes much longer to resolve if the object is not expected to be present in the situation and has not been explicitly introduced into the discourse before.

The conclusion which the psycholinguistic findings force one to make is that a theory which does not at least allow for a straightforward inclusion of background knowledge into the mental representations it postulates cannot be adequate.

It has to be admitted that one can always claim that one's theory is aimed at explaining only part of the representation which has been constructed in response to the actual linguistic input. According to this view, it is irrelevant that part of the representation in question has been constructed because of background knowledge. One is simply not interested in that part.

Although this reply might be satisfactory in its own right, it does not relieve one from the obligation to make clear at least how background knowledge could be integrated into the representations one has defined. Moreover, one has to explain what background information is activated, and how this is done. After all, it is certainly not the case that every piece of background information is relevant, as the following examples illustrate. Discourse (4) is consistent and readily understandable because we expect to find waiters in restaurants whereas discourse (5) is not because butchers are not expected to appear in restaurants [2].

(4) John went to a sea-food restaurant.
The waiter advised him to have the lobster.

(5) John went to a sea-food restaurant.
The butcher advised him to have the lobster.

Thus, only a limited amount of background information is activated at any

particular stage of a discourse, and, as the topic of interest is how people use language, this fact has to be explained.

3.1.5. Conclusion

Concluding then, we have mentioned four basic psychological facts which have to be taken into account in order for a theory of the human language processor to be satisfactory. Thus, such a theory is processor-centric, finite, allows for left-to-right processing, and takes the relevance of background information into account.

3.2. Linguistic criteria of adequacy

As pointed out, a satisfactory theory of the human language processor also has to take into account linguistic findings. This is true in particular of those theories dealing with the representations processors use when generating a piece of discourse or construct when comprehending a text. I want to mention two criteria specifically.

3.2.1. Empirical adequacy

The first linguistic criteria of adequacy is that the theory be empirically adequate in the sense that its predictions about how people understand a discourse agree with the "semantic" and "pragmatic" intuitions of native speakers. A theory which predicts counter-intuitive readings has to be rejected. In particular, if we restrict ourselves to the interpretation of pronouns, it is undeniable that often expressions or objects which at one stage of a discourse can be referred to using a pronoun or which at that stage of the discourse can be used for the interpretation of a given pronoun, can no longer play this role (much) later in the discourse. Now, if the theoretical entity which is postulated as a representation of the mental representation of discourse is to be used

in an account of pronoun use, then it will also have to be made clear how objects which at one stage of the discourse exists in this component can disappear out of it. An adequate theory not only has to account for the introduction of objects into the 'pronominalizable' part of the discourse representation, it also has to account for their disappearance out of it.

3.2.2. Non-adhocness

Another 'linguistic' requirement is that we want our theories to meet certain standards of 'formal' rigour. A theory which is completely *ad hoc* and does not specify precisely what possibilities are open to the language understander in response to a particular utterance in a particular discourse has to be rejected. I do not intend to imply that certain 'heuristics' do not play an important role but an appeal to 'heuristics' without some indications what these 'heuristics' look like is vacuous. The same applies of course to unmotivated appeals to 'pragmatics'.

3.2.3. Conclusion

Summarising then, a satisfactory theory about the human language processor has to meet two linguistic criteria of adequacy, namely consistency with the native speaker's "semantic" and "pragmatic" intuitions, and a certain formal rigour and non-adhocness.

4. Two caveats

Before I review the literature and discuss my own model, I want to make two caveats. The first concerns the type of language I will be discussing and the second the status of the discussion in general.

A number of authors have (for analytical reasons) made a distinction between two different functions of language and consequently two different types of language use. I am alluding to the distinction between "transactional" and "interactional" language use. (Brown and Yule,1983;1). Primarily transactional language is language which is used to convey "factual or propositional" information. The speaker primarily intends to efficiently transfer information to his hearer and it is essential that the hearer get the informative detail right. (Brown and Yule,1983;2). In primarily interactional language use, on the other hand, the speaker is concerned with the establishment and maintenance of social relationships. An example is discussion about the weather between people standing at a bus stop and waiting for a bus. (Brown and Yule,1983;3).

I do not want to make any claims about the relationship between these different types of language use, nor do I want to maintain that there is a sharp distinction between the two. As Brown and Yule (1983;1) stress, the distinction is one of analytical convenience and most real discourses have both interactional and transactional aspects. I will concentrate mainly on transactional language use. How far the system I present can be extended to also deal with interactional language use is an open question. I suspect that certain aspects of it can.

A second caveat concerns the status of the discussion in general. I will primarily discuss what the ideal speaker and the ideal hearer do in discourse. For instance, I will in general assume that both discourse participants are highly cooperative and sincere. The speaker will always be maximally informative and only try to transfer information which she considers to be true. The hearer on the other hand will always take the speaker to be authoritative and trustworthy.

It is clear that the speaker need not be maximally co-operative for an utterance to be successful. The reason for this is what one might call *the principle of retrospective updating of the knowledge base,* or the principle of retrospective updating for short. A number of rules governing the use of a particular type of expression state, as I will argue, that one cannot use a certain expression unless certain assumptions about the state of the knowledge base of one's hearer are warranted. Often, however, speakers seem to make the assumptions in question without any evidence. Clearly, communication does not always break down in these circumstances. The hearer, who as a somebody who knows the language, will also know the rules governing the use of the expression in question, will retrospectively update his knowledge base in response to the utterance in order to make the utterance which is technically speaking infelicitous, felicitous in the updated knowledge base. Thus, suppose that one can only utter an expression X if one can assume that the hearer has knowledge A activated. Now suppose that the speaker uses expression X even though the hearer has not activated knowledge A. Then technically speaking the utterance is infelicitous. But a co-operative hearer will realize that the utterance would have been felicitous if he had activated knowledge A. What the hearer may do in cases like these, is retrospectively update his knowledge base and activate A thus retrospectively making the utterance felicitous. Seuren (1985; 291) postulates a similar principle that is however more restricted in its application, the principle of *post hoc* or *backwards suppletion:-* if a definite term has no address, i.e. discourse object, to denote in D, i.e. the mental representation of the discourse, then such an address is created in response to a use of the definite term. In chapter 6. we will see that Clark and Marshall (1981) postulate a very similar principle.

5. Conclusion

In this chapter, I outlined a procedural view of language use according to which language cannot be studied without taking into account the structure of the human language processing mechanism. I also discussed a number of consequences of this view for theories about discourse comprehension. In the next chapter I will turn my attention to theories in the literature proposed to deal with discourse comprehension.

Footnotes:-

1. It is interesting that the thesis of Cognitive Similarity is also central in intuitionistic mathematics. Troelstra (1969,4), having said that the intuitionistic mathematician is mainly interested in constructions which exist in the mathematician's mind, writes:-

> The (mental) constructions we consider, are thought of as to exist in the mind of an individual (idealized) mathematician. The language of mathematics is an attempt (necessarily nearly always inadequate) to describe these mental constructions. Talking about intuitionistic mathematics is therefore a matter of suggesting analogous mental constructions to other people. Similarity between the thought processes of various human individuals makes such communication possible.

2. If one changes the tense in the second sentence of discourse (5) to a pluperfect, the discourse becomes acceptable and understandable. The event in which the butcher advised John to have a lobster no longer takes place in the restaurant but has taken place earlier (and presumably the butcher has been talked about earlier in the discourse or is mutually known for other reasons.) So, the activated restaurant script is not relevant for the interpretation of the definite description *the butcher*.

(1') John went to a sea-food restaurant.

The butcher had advised him to get the lobster.

Chapter 3:- Theories of the human language processor

1. Introduction

In this chapter, I want to discuss theories which have been proposed to account for discourse production and discourse comprehension. As stated in the previous chapter, satisfactory theories fall apart in two parts:- theories about the overall structure of the human language processor, and theories about the specific representations used in discourse production and discourse comprehension. Unfortunately, very few authors have actually paid attention to both of these separate aspects. One of the few exceptions is Johnson-Laird who formulates proposals both about the overall architecture of the processor and about the detailed representations used in discourse production and comprehension. Given the need for two separate parts in a satisfactory theory, I will divide the literature review in two sections. First, I will discuss theories about the overall structure of the human language processor. Secondly, I will discuss theories about the specific representations used in discourse production, and constructed in discourse comprehension. But first I want to discuss another approach to theories about the human language processor, which can best be described as psychologizing the formal approach to language.

2. Psychologizing the formal approach

At the beginning of a paper which discusses the relationship between psychological semantics and truth conditional or model-theoretic semantics, Johnson-Laird (1982,1) writes :-

> Logicians have only related language to models in various ways; psychologists have only related it to the mind; the real task, however, is to show how language relates to the world through the agency of the mind.

Given the sub-division of the relation between the world and language into two sub-relations, one between the world and the mind, and the other between the mind and language, it is tempting to develop a theory of the human language processor which is analogical to the system in Montague (1973). Montague defined an algorithm to map syntactic analyses of sentences of a fragment of English into formulae of a logical language. The logical language whose formulae Montague translated the natural language sentences into, had been shown to have a clear and unproblematic model theoretic interpretation [1]. Given the proof in Montague (1970b) that the translation algorithm, i.e. the mapping from the syntactic analyses of the English sentences onto formulae of the logical language, preserved the model-theoretic interpretation, one thus had an (indirect) model-theoretic interpretation for the English sentences.

In the psychological reconstruction of the Montague programme, one could regard the intermediate language as a language of thought (Cf. Fodor, 1976), which is then mapped into a model, or the world. One thus replaces Montague's logical language by a mental language but keeps the rest of the programme unchanged. The usefulness of this programme then partly depends on the independent use one can find for this intermediate level.

The idea of psychologizing Montague's work goes against the spirit of his work. In the first place, Montague was not interested in psychology at all. He saw English as just another formal language, perhaps more complicated than but not theoretically different from the language of first-order predicate calculus. He wrote (Montague, 1970a):-

I reject the contention that an important theoretical difference exists between formal and natural languages.

Secondly, Montague regarded the intermediate level as not necessary, but purely one of convenience. Montague (1973) included it, but Montague (1970a) had a direct model-theoretic interpretation of English, i.e. a model-theoretic interpretation without a mediating translation of the English sentences into formulae of a logical language.

However, the fact that the psychologically oriented analog of Montague's programme goes against the spirit of his work, has not deterred psychologists and formal semanticists. Indeed, Kintsch (1974) and Fodor (1976) propose to see the mental representation of a text as a set of propositions in some mental language.

Johnson-Laird (1982) points out that the research strategy of "psychologising" Montague's programme crucially depends on the assumption that the relation between language and the mental representations of sentences or texts can be described independently of the relation they bear to the world. In order for the strategy to be viable it has to be possible to get from discourses to mental representations of discourse without taking into account the world as such, or the world as seen by the discourse participants. One can even go further and say that the research strategy depends on the assumption that the relation between language and the mental representations of sentences can be described independently of factors other than the linguistic input. If the transition from actual sentences to mental representations of sentences is more or less like Montague's translation algorithm, then this has to be describable independently of factors such as the mental representation of the immediate environment in which the sentence is being used, knowledge of the world, previous discourse etc.

This assumption however is unwarranted. First, the inputs to Montague's translation algorithm are syntactic analyses of sentences rather than just sentences. As a consequence, his algorithm gives different results for the different readings of syntactically ambiguous sentences. Since Montague's work is based on what Bach calls the rule-to-rule hypothesis (for every syntactic rule there is one and only one semantic rule), for semantically ambiguous sentence, there is a separate syntactic analysis as well. The problem for the psychologically oriented theorist is clear:- the utterances hearers or readers are faced with are not syntactically disambiguated. Now, one of the requirements one has to put on the propositional representations is that they are in general unambiguous, especially if it can be shown that the various purported readings have different effects on the following discourse. If a sentence has two readings, the theory should generate two different propositional representations. The transition from sentences to mental representations therefore has to be disambiguating. But the disambiguation of an utterance depends on factors such as previous discourse and real-world knowledge (Altmann, forthcoming). As a consequence, the transition from utterances cannot be described on the basis of the linguistic input alone.

Another minimal requirement of adequacy on the mental representations of sentences is that intersentential pronouns are given their right interpretation. Again, it is known that the resolution of anaphora in general, and intersentential anaphora in particular, depends often on real world knowledge and previous discourse.

A third argument is demonstrative pronouns. An adequate representation of discourse will have to make clear what the hearer takes the speaker to deictically refer to. This obviously cannot be done if one does not relate the mental representation of utterances to the world, or to the mental representations of the

world.

It follows that a mental representation of discourse based purely on the linguistic information in the utterance is not satisfactory. The mental representation of discourse is also constructed on the basis of the hearer's mental representation of the environment in which the sentences is being used, his world-knowledge, his recognition of the speaker's intention, his understanding of previous discourse etc. Using language amounts to constructing mental representations based not only on the linguistic input but also on other factors such as the afore-mentioned, and thus this simplistic approach to 'psychologizing' the Montague programme fails [2]. I will return to this approach to the mental representation of discourse when I discuss the truth-conditional approach in section 4.2. of this chapter.

3. Theories about the general structure of the processor

In this section I want to discuss two theories which have been proposed as models for the architecture of the human language processor. The first model is that of Sanford and Garrod and bears a large resemblance to the model I will present in chapter 4. The second model I want to discuss in some detail is that of Johnson-Laird (1983, chapter 11).

3.1. Sanford and Garrod's model

One of the best known models of the human language processor is the model of Sanford and Garrod. The model has been developed to account for the understanding of written language.

Sanford and Garrod (1981;157-60) claim that, from the point of view of a semantic processing system, it is useful to see the various elements in a sentence as instructions to the hearer/reader to perform certain operations. When the hearer encounters a verb in a sentence for example, he has to retrieve the semantic representation of the verb. But understanding a piece of discourse also involves constructing a mental representation of it. The hearer has to keep a record of the information he received from the speaker. The comprehension process thus consists of both retrieving the appropriate information and constructing a representation of the text, partly based on the retrieved information.

Both the retrieval and the construction process can be specified in terms of three variables. For the retrieval process, they are:-

(i) the memory domain to be searched,
(ii) a partial description of the information the processor is looking for,
(iii) the type of information to be retrieved.

The construction process can be characterised in terms of

(i) the memory domain in which the construction is to be recorded,
(ii) a description of the information to be incorporated,
(iii) the type of structure to result

Sanford and Garrod distinguish between four components in the model for the language understander:- Explicit Focus, Implicit Focus, Long-Term Semantic Memory and Long-Term Text Memory. The different components are to be seen as memory partitions, i.e.

> independently addressable and capable of being treated by the processor as a distinct search domain.
>
> Sanford and Garrod (1981,158).

Explicit focus is the memory partition that contains representations of entities and events explicitly introduced in the text. It is a short-term store of limited

capacity. In Sanford and Garrod's model tokens representing entities explicitly introduced into the discourse, point to computational spaces in Explicit Focus. The size of the space is the realization of the degree to which it is foregrounded, or activated. The larger the computational space a token points to, the more foregrounded the corresponding entity is, and the more likely the speaker is to use a pronoun to refer to it. One could in fact "measure" the degree of foregrounding in terms of the ease with which one can "access" it using a pronoun. As discourse goes on and new entities are introduced, the highly foregrounded tokens will gradually return to the background and eventually disappear out of Explicit Focus altogether.

Implicit Focus is the component which contains the background knowledge, scenarios etc., relevant for the comprehension of the utterance currently being processed. Sanford and Garrod (1981,162) say that it can be thought of as

> a partition of long-term memory which is simply currently priviliged in terms of ease of access.

The knowledge store out of which information in Implicit Focus has been activated is called Long-Term Semantic Memory. Long-Term Semantic Memory thus contains the long-term knowledge a discourse participant brings to a discourse.

The last component, Long-Term Text Memory, contains a long-term representation of the content of the text. The reason for distinguishing between this component and Long-Term Semantic Memory is that Sanford and Garrod claim that it is important to retain a separation between memory for the text itself and other general knowledge.

The definition of a memory partition entails that in order to have sufficient reason for distinguishing between components in memory, Sanford and Garrod have to show that two components are indeed independently addressable, or are treated by

processors as distinct search domain. Given their theoretical assumptions about the comprehension process, there are two ways of doing this. In the first place, they can show that a particular type of linguistic expression is an instruction to retrieve information from one of the components but not the other. The second way would be to show that the result of the construction process has to be recorded in one particular component and not the other. A theory-independent way of testing this would be to show that subsequent retrieval of the discourse can only be viewed as a retrieval of information in one of the components.

If we apply this to the distinction between Implicit Focus and Long-Term Text Memory, either Sanford and Garrod have to show that the retrieval procedures for a particular type of linguistic expression have as their search domain Implicit Focus but not Long-Term Text Memory or vice versa, or they have to show that subsequent retrieval can be explained solely in terms of the contents of Long-Term Text Memory or exclusively in terms of what is in Implicit Focus. I will argue that neither of these criteria licenses a distinction between these two components.

Sanford and Garrod argue for the distinction between Implicit Focus and Long-Term Text Memory on the basis of an analysis of definite descriptions. They claim that the retrieval procedures for definite descriptions specify that the memory domain to be searched is focus, i.e. implicit focus or explicit focus [3]. A search triggered by a definite description is successful in Explicit Focus if one can find a token which has the property described by the noun in the definite description; if the search domain is Implicit Focus, then one has to be find an entity whose existence is implied by the background knowledge currently in implicit focus which has the property in question.

Example (1), taken from the novel by Jean Rhys *After leaving Mr Mackenzie*, shows that definite descriptions can also be used to bring back into focus entities which were in focus earlier in the discourse. They not only set up entities which act as slot-fillers in the current frame, they also re-activate entities which have been spoken about before.

> (1) The bed was large and comfortable, covered with an imitation satin quilt of faded pink. There was a wardrobe without a looking glass and a red plush sofa and - opposite the bed and reflecting it - a very spotted mirror in a gilt frame.
>
> The ledge under the mirror was strewn with Julia's toilet things - an untidy assortment of boxes of rouge, powder, and make-up for the eyes. At the farther end of it stood an unframed oil-painting of a half empty bottle of red wine, a knife, and a piece of Gruyere cheese, signed 'J. Grykho,1923'. It has probably been left in payment of a debt.
>
> Every object in the picture was slightly distorted and full of obscure meaning. Lying in bed, Julia would sometimes think: "I wonder if that picture's any good. It might be; it might be very good for all I know .... I bet it is very good too."
>
> But really she hated the picture. It shared, with the colour of the plush sofa, a certain depressing quality. The picture and the sofa were linked in her mind. ...'

The definite description *the plush sofa* in the last paragraph of the quote has the function to bring back into focus an object which has been introduced earlier in the discourse, and whose existence does not depend directly on general knowledge one has about the setting of the scene. So, definite descriptions trigger searches not only of Implicit Focus but also of other memory partitions. The question now arises whether this other memory partition is Explicit Focus or not. If it is, then Sanford and Garrod are home and dry; if not, and if moreover the other partition can be shown to be Long-Term Text Memory, then they are in trouble, as there would be no linguistic reason to distinguish between these two components.

It is clear from example (1) that definite descriptions can be used to re-introduce entities which have been introduced in the text before and must have existed in Explicit Focus. If one is to believe Sanford and Garrod's analysis of definite

descriptions, and sees them as instructions to search focus, then one is committed to the view that the objects remain in Explicit Focus almost indefinitely. However, Sanford and Garrod (1981;162) write:-

> If, as we suggest, explicit focus is a short-term store of limited capacity, then the amount of information being held in it must, of course, be limited. Accordingly, we suggest that explicit focus only contains tokens and scenario pointers. As new tokens are added, old ones will gradually diminish in terms of the computational space to which they point, until, eventually, they are no longer in focus at all.

Of course, things which disappear out of Explicit Focus cannot go into Implicit Focus, since this component contains an activated part of Long-Term Semantic Memory. Given the role of Long-Term Text Memory, it is more likely that the entities disappear into this partition. Thus, definite descriptions can be used to instruct the reader to access material either in Explicit Focus, or Implicit Focus, or Long-Term Text Memory. It follows that one cannot use definite descriptions as evidence that Implicit Focus and Long-Term Text Memory are independently addressable.

Sanford and Garrod do discuss the possibility of bringing back into focus earlier topics. To this end, they introduce the notion of 'secondary processing'. They write (Sanford and Garrod, 1981;167):-

> referential resolution can be successful because there is either a token in explicit focus, or a suitable slot in implicit focus, or both. In any piece of discourse, this need not be the case, and when such resolution is not possible *secondary processing* is called for.

Secondary processing is done over an extended search domain. The appropriate search domain now also includes the long-term memory partitions, Long-Term Text and Long-Term Semantic Memory. Since the search domains are larger than in the case of primary processing,

> secondary processing is both slower than is primary, and longer partial descriptions are required for its success. Taking a general view of comprehension once more, secondary processing provides the means by which new topics can be introduced, or earlier topics reintroduced to current focus.
>
> Sanford and Garrod (1981,171).

There are various problems with this notion of secondary processing. In the first place, its introduction in the theory has the consequence that one can no longer describe the retrieval procedures for definite descriptions in terms of the three variables mentioned earlier in this section. It will be remembered that Sanford and Garrod described the search domain for the retrieval procedures for definite descriptions as focus, either Explicit or Implicit Focus. The fact that there is also secondary processing means that this no longer is true.

Sanford and Garrod give two characteristics for secondary processing. First, secondary processing is supposed to take more time, and secondly it is supposed to require longer partial description to guide the search.

The second criterion can easily be dismissed. In some cases, we need relatively long partial descriptions even if the search is confined to focus, e.g. in a discourse such as (2).

> (2) There are three blocks, a red one and two black ones. One of the black ones is small and the other is large. The red block is a bit smaller than the large black one. The little black block is on top of the red one.

Note that in the last sentence of this constructed example one had to construct a definite description using all the information one had about the object one was talking about in order to be sure that the hearer would have no doubt which object to access. Thus, we have an example of an extremely long partial description which has to be used even though the intended referent is in Explicit Focus. Conversely, a minimal description such as *the sofa* would have been sufficient in

discourse (1), even though the intended referent must exist in Long-Term Text Memory. Similarly, one can come up with examples where an object is introduced with a minimal description. For instance, discourse (3) which is the first sentence of the Hemingway story *Cat in the rain*.

> (3) There were only two Americans staying at the hotel. They did not know any of the people they passed on the stairs on their way to and from their room.

In this example, the definite description *the hotel* introduces hotel-knowledge into the discourse. The definite descriptions *the stairs* and *their rooms* and the state of affairs described in the second sentence clearly have to be understood against the background of a hotel. Hemingway thus uses a minimal definite description to introduce a frame.

The other criterion, that of slower processing, needs to be tested. It may be true that definite descriptions which require secondary processing require more processing time. The question however remains what the exact explanation of this phenomenon would be. It is unlikely that the most feasible explanation would be in terms of the size of the search domain. For if this were the explanation, then it would be predicted that the interpretation of indefinite NPs, which presumably have as their search domain Implicit Focus and Long-Term Semantic Memory (and maybe even Long-Term Text Memory), would be considerably slower than the interpretation of any other type of expression.

But even if the notion of secondary processing can be shown to be of relevance to a theory of language processing, then it still remains to be shown that one needs a distinction between Implicit Focus and Long-Term Text Memory, as Sanford and Garrod maintain. At least from the point of view of search procedures, it

seems, there is little if any reason to conclude that the processor treats these components as distinct search domains.

It must be noted that Sanford and Garrod could still use the argument that Implicit Focus and not Long-Term Text Memory, or vice versa, contains all the information which people use when asked to recall a piece of text. However, given the data on text-recall given in chapter 2. and indeed Sanford and Garrod's discussion of the representation of text in memory, it is highly unlikely that they would be willing to uphold such a claim.

The conclusion then has to be that there is little reason to distinguish between Implicit Focus and Long-Term Text Memory, as Sanford and Garrod do, and that there is reason to believe that there is a "component" which simultaneously plays the different roles Sanford and Garrod ascribe to Implicit Focus and Long-Term Text Memory.

3.2. Johnson-Laird's procedural model

Although Johnson-Laird is probably best known for his detailed proposals concerning the structure of mental models, which will be discussed in section 4.4. of this chapter, chapter 11 of his book *Mental Models* contains a proposal for an overall theory of comprehension. Johnson-Laird first outlines a number of assumptions and then gives a list of procedures which the processor is claimed to use when interpreting a discourse. He does not make any detailed proposals for discourse production. I will first discuss the assumptions underlying Johnson-Laird's approach, and then turn to the procedures which he proposes people use when comprehending a discourse.

Johnson-Laird (1983,246-37) lists five assumptions underlying his theory of discourse comprehension. They are:-

1. The processes by which fictitious discourse is understood are not essentially different from those that occur with true assertions.

2. In understanding a discourse you construct a single model of it.

3. The interpretation of discourse depends on both the model and the processes that construct, extend and evaluate it.

4. The functions that construct, extend, evaluate, and revise mental models, unlike the interpretations functions of model-theoretic semantics, cannot be treated in an abstract way.

5. A discourse is true if it has at least one mental model that can be embedded in the model corresponding to the world.

According to Johnson-Laird (1983,249-50), the process of translating an assertion into a mental model requires several general procedures. They are:-

1. A procedure that begins the construction of a new mental model whenever an assertion makes no reference, either explicitly or implicitly, to any entity in the current model of the discourse.

2. A procedure which, if at least one entity referred to in the assertion is represented in the current model, adds the other entities, properties, or relations, to the model in an appropriate way.

3. A procedure that integrates two hitherto separate models if an assertion interrelates entities in them.

4. A procedure, the verification procedure, which, if all entities referred to in an assertion are represented in the current model, verifies whether the asserted properties or relations hold in the model.

5. A procedure that adds the property or relation (ascribed in the assertion) to the model in the appropriate way. This procedure is used when the verification procedure does not return a definite truth value for the assertion.

6. A procedure, called whenever the verification procedure returns the value "true", that checks whether the model can be modified in a way that is consistent with the previous assertions but so as to render the current assertion false.

7. A procedure, called whenever the verification procedure returns the value "false", that checks whether the model can be modified in a way that is consistent with the previous assertions but so as to render the current

assertion true.

Johnson-Laird (1983,244) assumes that the process of discourse understanding has two stages. One first constructs a propositional representation of the utterance, which is very near the surface form of the sentence to be understood, and is the result of a first superficial understanding. In a second, and optional, stage the propositional representation serves as a partial basis for the construction of a mental model. It triggers one of the procedures listed above. Which of the procedures is triggered is a function of a variety of factors including the referring expressions in the propositional representation, the context as represented in the current mental model, and the background knowledge that is triggered by the sentence.

There are a number of problems with the proposed procedures, and especially with the last two procedures. Johnson-Laird probably includes procedures 6. and 7. because most of his work on mental models deals with the problem of spatial inference (Mani and Johnson-Laird, 1982) and the problem of solving syllogisms (Johnson-Laird and Steedman, 1978; Johnson-Laird, 1983;64-145). His theory on spatial and syllogistic reasoning can be summarized as follows. When solving problems of this sort people first construct a mental model on the basis of the first premise, and then do the same on the basis of the second premise. They then combine the two premises, and read off a (preliminary) conclusion from this model. Finally, they try to modify the model in order to make the conclusion false while leaving the truth of the premises intact. If they fail to do so, then the conclusion is put forward as a valid conclusion; if they succeed and can modify the model in such a way that the first tentative conclusion is falsified, then they read off another preliminary conclusion and try to falsify this one. Johnson-Laird (implicitly) claims that similar processes are going on in discourse comprehension.

There are various reasons to doubt the truth of this claim. If we initially restrict our attention to syllogistic reasoning then the first problem which crops up concerns the time it takes people to solve a syllogism. In general, it takes 5 to 30 seconds to solve a syllogism and a syllogism involves only two premises. Assuming that not too much time is taken up by the actual formulation of the conclusion, it should take about as much time to process a number of sentences in natural language. However, it is clear that understanding discourse happens at a faster rate than 2 or 3 sentences per 20 seconds.

To be fair to Johnson-Laird one has to admit that it may take longer to translate into a mental model sentences used in syllogisms than it takes to translate other sentences, especially given the necessity for producing an explicit response in the syllogism task, a necessity not present in ongoing discourse processing. On the other hand, given that there can be quite a number of sentences in any single discourse, the complexity of the model would increase considerably, thereby making manipulation of the model presumably harder and more time-consuming.

A second criticism concerns the way the procedures in question are supposed to work. If the current assertion is found to already have a definite truth value in the model, the procedures try to modify the model so that it is still consistent with the previous sentences in the discourse while giving the current assertion a different value in the model. Since mental models are constructed on the basis of propositional representations, it follows that for the procedures in question to work, the propositional representations have to remain available. One has to check the proposed alterations of the model against the stored propositional representations to make sure that they are not directly ruled out by one of the premises [5]. As Johnson-Laird (1983,254) himself points out, it follows that the propositional representations of

sentences uttered earlier in the discourse remain available [6].

The problem is that there is ample evidence that people do not normally store texts using propositional representations which are very near the surface form of the sentences contained in the discourse, some of which was presented in the previous chapter. People build mental representations of texts which, whatever form they take, are quite remote from the surface forms of the sentences they received, and include a large amount of the relevant background knowledge.

Johnson-Laird (1983,160-2) discusses an experiment by Mani and himself (Mani and Johnson-Laird, 1982) which is intended to support the view that there are different levels of representations. They presented subjects with spatial descriptions which were either determinate in that they allowed for only one spatial lay-out, or indeterminate in that they allowed for more than one possible spatial lay-out. It emerged that subjects remembered the gist of determinate descriptions much better than that of the indeterminate descriptions, in the sense that sentences from the text or inferrable sentences were ranked prior to confusion items significantly better for the determinate descriptions than for the indeterminate ones. On the other hand, the verbatim details of indeterminate descriptions were remembered significantly better than the verbatim details of the determinate descriptions. Mani and Johnson-Laird concluded from these findings that there are two levels of representation: a propositional level and a mental model level.

However, it still does not follow that in every discourse propositional representations are mentally stored. What the Mani and Johnson-Laird experiment shows is that there are two forms of representation of text in memory:- texts can be stored either in a propositional representation or in a representation of the mental model type. What Johnson-Laird needs to show in order for his theory of

comprehension to go through is that representations of the propositional type and representations of the mental model type are used simultaneously and throughout the discourse. What he has shown is that there are two types of representation; what he needs to show is that the propositional representation and the mental model co-exist throughout the discourse. After all, the testing procedures 6. and 7. imply that one compares alternative mental models, i.e. representations of the mental model type, with propositional representations of previous sentences. The experiment has not shown that one stores mental models and propositional representations simultaneously, and indeed strongly suggests the opposite conclusion.

The criticism in the previous paragraphs is not intended as an argument against using both propositional representations and mental models. There are many arguments for believing in a propositional representation mediating between the spoken string and the mental model. Apart from the Johnson-Laird and Mani experiment, Johnson-Laird (1983;394) also cites the evidence of VP-deletion, which can only be explained on the basis of the linguistic string. What the criticism is aimed against is the fact that for the procedures to be applicable the propositional representation has to remain available all the time, and that is psychologically implausible.

From a discourse comprehension point of view, most of Johnson-Laird's experiments on spatial and syllogistic inference are artificial. The reason is that, unlike in spoken language, the premises remain available while the subject is reasoning. Subjects therefore do not need to store mentally propositional representations of the premises. They can concentrate completely on the reasoning task which, as Johnson-Laird has quite convincingly shown, requires mental representations of the model type. If the mental model has to be revised then subjects

can compare their new mental model with the information in the sentences they have before them. Subjects who find out that it is impossible to univocally construct a mental model on the basis of the sentences given, then have a choice of storing the propositional representations, or to make certain decisions so as to make the construction of a single mental model possible. Johnson-Laird (1983;164) mentions an experiment by Stenning (1981) who found that subjects sometimes do take the latter option and construct a mental model by arbitrarily adding information which cannot be retrieved or deduced from the sentences presented to them.

The conclusion then has to be that from a purely theoretical point of view, the procedures which Johnson-Laird proposes to account for discourse comprehension are somewhat dubious, because they rely on storage of propositional representations alongside and simultaneous with mental representations of a model type. But in practice, this problem is less critical. Radical revision of a mental model built on the basis of a discourse is not often called for. Speakers and hearers generally obey Gricean principles, and make discourses as easy to comprehend as possible, and therefore, from a practical point of view, it is not necessary to keep both representations around (Cf footnote 6).

4. Theories about the mental representations of discourse

There are a number of different formal proposals in the literature about the specific mental representations of discourse. They can be divided roughly in three different groups. First, there is the "model-theoretic" group which postulates some theoretical entity corresponding to the mental representation of a discourse which is interpreted through some embedding function into standard logical models. In a sense, this approach differs least from the normal model theoretic account of formal semantics as proposed by Montague. Logical models still play a major role in the interpretation of

a discourse and a sentence. The only difference is that there is an intermediate level between the language and the models and this intermediate level is a formal representation of the mental representation people construct in response to a discourse. Examples of this approach are Kamp (1981) and Heim (1982).

The second approach can be called the "description" approach. It has been developed mainly to account for the interpretation of pronouns. The idea is to generate descriptions of discourse objects on the basis of a given sentence and then use them in the interpretation of anaphora. Webber (1979) and Richards (1984) are examples.

The final approach is that of mental models. In response to the discourse a hearer constructs a mental model, which more or less takes over the role played by the mathematical model in formal semantics.

In each of the cases discussed below I interpret the theory as a theory of the mental representation of discourse, whether or not the author chooses to see it so. This will lead in a number of cases to criticisms which are not entirely fair. Thus, I will judge a number of theories which are proposed mainly for semantic reasons on psychological criteria. Although in the end I agree with Kamp (personal communication) that this is probably the right way to go, I realize that not everybody shares this view. What I criticise in this chapter are therefore often my interpretations of the various systems proposed in the literature.

Before I turn to a discussion of the various proposals for the representation of discourse, I would first like to discuss briefly the so-called donkey sentences. The reason is that they have taken a central place in discussions about the discourse representation. Indeed, part of the reason why formal semanticists became interested

in discourse representations was the fact that Kamp (1982) claimed to be able to solve the problem of these sentences by means of his Discourse Representation Structures.

4.1. Digression:- donkey sentences

Given the importance of the so-called donkey sentences for both Kamp's and Richards' theory, it is necessary to divert our attention to sentences of this type for a moment. Examples of sentences of this type are (3) and (4).

(3) Every miner who owns a donkey beats it.

(4) If a miner owns a donkey he beats it.

Formal semanticist have traditionally analysed indefinite NPs as existential quantifiers. Although this analysis is not without problems, I will assume for the sake of argument that the analysis is correct and that the indefinite NPs *a miner* and *a donkey* in (4) have to be analysed as involving an existential quantifier. The problem posed by the donkey sentences is that translating the indefinite as an existential quantifier would lead to binding problems. That is, the most straightforward translation of (3) would be (5) and (5) would be true if there was a miner who did not own a donkey, or even if there was an object which was not a miner or a donkey, clearly a counter-intuitive result.

(5) (Ex)(Ey)[(miner(x) & donkey(y) & own(x,y))
--> beat(x,y)]

Moreover, it was also argued that the only way to capture the truth conditions of (4) was by giving the indefinites *a miner* and *a donkey* universal import, i.e. translate them as universally quantified NPs, as in (6).

(6) (x)(y)[(miner(x) & donkey(y) & own(x,y))
--> beat(x,y)]

There certainly is no uniform agreement about whether (6) is the correct

rendering of the truth conditions of the sentences in question, as Kamp (1981,179) himself admits. For some speakers the indefinites *a miner* and *a donkey* should be given 'generic' rather than 'universal' import. The claim made concerns the prototypical miner and the prototypical donkey which he owns, rather than all individual miners and all individual donkeys which they own (Cf. Seuren, 1985; section 4.2.4). Thus, the sentence in question is almost equivalent to (7) [7].

(7) A miner who owns a donkey beats it.

It is in general not very useful to discuss varying intuitions to any great length. I will therefore accept the claim that indefinites occurring in antecedents of conditionals or in relative clauses governed by a universally quantified NP have to be translated as universal quantifiers. It is sufficient to say that the intuitions are not as clear-cut as one would like them to be. In section 4.3 of chapter 6 I will return to donkey sentences.

4.2. The model theoretic approach

Proponents of the model-theoretic approach to the description of the mental representation of discourse postulate a structure which intermediates between language and the world. One can graphically represent the model-theoretic approach as follows:-

discourse --> discourse representation --> world

Kamp (1981) calls the intermediate structures *Discourse Representation Structures* whereas Heim (1981) uses the term *file*.

Kamp writes about Discourse Representation Structures

> I conjecture that the structures which speakers of a language can be non-trivially described as forming to represent verbal contents are, if not formally identical, then at least very similar to the representations here defined.
>
> Kamp (1981,282).

Kamp thus claims that the formal structures the theory postulates have to be regarded as formal representations of the mental representation of discourse. Heim (1982) is less explicit about the status of her files, but by comparing entries in the file, or file cards, to Karttunen's discourse referents (Karttunen, 1976) and Webber's discourse entities (Webber, 1979), and files themselves to Stalnaker's notion of common ground (Stalnaker, 1979), she gives her notion of file a strong cognitive flavour.

One of the central aspects of the model-theoretic approach to the mental representation of discourse is the central role it gives to the standard models from logic. The intermediate structures which are postulated get their semantic status because of the relationship which is defined between them and the standard models.

Apart from the theoretical reasons of giving a formal description of the mental representation of discourse, Kamp and Heim also see more linguistic reasons for their respective systems. Kamp uses his system to give an account for the donkey sentences, whereas Heim uses hers to account for the indefiniteness-definiteness distinction. Rather than discussing both systems in full detail I will discuss only Kamp's system. The main reason is that I believe that all the criticism which I raise against Kamp also apply in some form to Heim's system. Given that Kamp's system has the advantage of greater formal rigour, I will restrict myself to Kamp's system.

4.2.1. Kamp's theory

Kamp's theory consists of two parts. The first part is a formal system for constructing

the representations which users of the language are claimed to form in response to verbal input, so-called *Discourse Representations Structures.* They are partially ordered sets of *Discourse Representations* which in turn are ordered pairs consisting of a set of individual constants, the discourse referents, and (occurrences of) formulas of a language closely resembling a quantifier-free first-order language. The second part of the theory links these representations to arbitrary (standard logical) models, thus defining truth conditions. The basic idea is that a discourse D is true in a model M if and only if the Discourse Representations Structure DRS constructed on the basis of D is compatible with M. I will return to some of the details of the formal definition of compatibility later [8].

4.2.2. Constructing Discourse Representation Structures

Intuitively, one constructs a Discourse Representation Structure as follows:- given a discourse D in a language L one constructs the first Discourse Representation whose set of individual terms is originally empty and whose set of formulas originally contains just the first sentence in the discourse. One then applies the rules for the construction of Discourse Representation Structures, as far as possible. Some rules add new discourse referents and new sentences to a Discourse Representation. Other rules also add more Discourse Representations to the Discourse Representation Structure. It is important to point out that the order in which the rules are applied to a given sentence depends on the syntactic analysis of the sentence.

After the rules, when applicable, have been applied to the first sentence of the discourse, one adds the second sentence of the discourse to the principal Discourse Representation, i.e. the highest Discourse Representation, which still contains the first sentence in the discourse. One then applies all possible rules for Discourse Representation Structure construction. Then one treats the next sentence in the

discourse in the same way, and so on until the last sentence in the discourse has been dealt with.

The partial ordering of the Discourse Representations in a Discourse Representation Structure is a consequence of the fact that the construction rule triggered by conditionals and universal sentences instruct one to create two new Discourse Representations in the Discourse Representation Structure. These new Discourse Representations are called subordinate to the Discourse Representation which contains the occurrence of the universal or conditional sentence.

I will illustrate Kamp's theory by working through an example in some detail. Consider sentence (8).

(8) If Pedro owns a donkey he beats it.

The first step in the construction of a Discourse Representation Structure for this sentence is to create a Discourse Representation with (8) as its only element.

$$\text{If Pedro owns a donkey he beats it}^{m1}$$

The next step is determined by the fact that sentence (8) is a conditional. We thus have to apply the rule for conditional sentences to m1. The rule tells us to create two Discourse Representations which are subordinate to m1, containing respectively the antecedent and the consequent of the original sentence.

$$\text{If Pedro owns a donkey he beats it}^{m1}$$

$$\text{Pedro owns a donkey}^{m2} \qquad \text{he beats it}^{m3}$$

We now have to apply the appropriate construction rules to m2, and the first rule we can apply is that governing proper names. It says to create a new discourse

referent in the principal Discourse Representation and to add a sentence to it asserting the identity between the newly created discourse referent and the name. Moreover, one has to add a sentence to the Discourse Representation the sentence which gave rise to an application of this construction rule occurred in. It is formed from the original sentence by substituting the discourse referent for the proper name.

```
                                                        m1
                          .
                          u
                    u = Pedro
   If Pedro owns a donkey he beats it


                          m2                        m3
Pedro owns a donkey                he beats it
  u owns a donkey
```

The next rule to be applied is triggered by the indefinite NP *a donkey*. It instructs one to introduce a new discourse referent into the Discourse Representation which contains the sentence with the indefinite NP in question. Moreover, one has to add two new sentences. The new sentences are constructed by predicating the noun in the indefinite NP of the newly created discourse referent, and by substituting the discourse referent for the indefinite in the original sentence.

```
                                                m1
                    .
                    u
              u = Pedro
   If Pedro owns a donkey he beats it


                          m2                        m3
          .                                he beats it
          v
Pedro owns a donkey
  u owns a donkey
   donkey(v)
   u owns v
```

This is the last step one can take as far as Discourse Representation m2 is

concerned. However, there is one rule which as to be applied twice to Discourse Representation m3. This rule, which deals with pronouns, instructs one to find a suitable antecedent for the pronoun, i.e. a suitable discourse referent in a superordinate Discourse Representation, and add a new sentence to the Discourse Representation in which the sentence with the pronoun occurred. The new sentence is formed from the old one by substituting the discourse referent for the pronoun. This rule is the only one which relies on the partial ordering of the Discourse Representation in a Discourse Representation Structure:- one condition on what a suitable antecedent for a pronoun is is that the discourse referent exists in a superordinate Discourse Representation. The results of applying this rule twice to the consequent of the sentence we are dealing with, taking *Pedro* as the 'antecedent' of *he* and *a donkey* that of *it* is as follows:-

```
                                                        m1

                          .
                          u
                    u = Pedro
   If Pedro owns a donkey he beats it


                          m2                            m3
            .                              he beats it
            v                               u beats it
Pedro owns a donkey                         u beats v
   u owns a donkey
     donkey(v)
     u owns v
```

Kamp gives just one other construction rule for universal sentences. The idea is to create two subordinate Discourse Representation, one containing a sentence based on the subject of the sentence and one containing a sentence based on the predication made. Thus in response to sentence (9) one would construct two new Discourse Representations, the first one originally containing just a discourse referent x and the sentences 'man(x)' and 'x owns a donkey', and the second containing only

the sentence 'x beats it'. As a consequence, sentence (9) and (10) give Discourse Representation Structures which only differ in the sentence contained in the principal Discourse Representation.

(9) Every man who owns a donkey beats it.

(10) If a man owns a donkey he beats it.

Kamp has two reasons for wanting his subordinate Discourse Representations. They not only determine the set of possible antecedents of a given pronoun, but they also enable him to give a uniform account of indefinite NPs including those occurring in donkey sentences. In order to see how this works we have to turn to the truth definition.

4.2.3. The truth definition

The basic idea underlying the truth definition of a universal or conditional sentence is as follows:- A universal or conditional sentence is true in a model M if and only if every verification of the first Discourse Representation constructed in response to the sentence can be extended to a verification of the second Discourse Representation. It is this idea which gives indefinites when they occur in a 'donkey-context', i.e. either in the antecedent of a conditional or in a relative clause hanging of a universally quantified NP, their universal import.

We can illustrate this if we look at the Discourse Representation Structure which we constructed in response to sentence (8). This sentence if true according to the truth definition if every verification of the first Discourse Representation can be extended to one of the second. Thus, if there is a donkey which Pedro owns, then this donkey will also have to be beaten by Pedro. However, in order for the sentence to be true, the above has to hold for every verification of the first Discourse Representation, i.e. every donkey which Pedro owns also has to be beaten by Pedro. We thus see how

the indefinite NP *a donkey* gets universal import in sentence (8).

4.2.4. An appraisal of Kamp's theory

Kamp's theory suffers from a number of shortcomings. As a linguistic theory, Kamp's theory suffers from the fact that he does not account for the fact that entities which can be referred to by using a pronoun at one stage of the discourse can lose this status later in the discourse. Kamp (1981,283) says that he is not concerned with the strategies used in selecting the referents of personal pronouns but rather with the sets of referential candidates from which the strategies select, and he could of course reply to the above objection that the disappearance of referential candidates out of these sets is a process which has to do with the selecting strategies. However, as this implies that in principle something remains for ever available for pronominalization, this reply has to be rejected.

One of the main achievements Kamp would claim for his theory is the fact that he is able to provide a uniform analysis of indefinite NPs across a variety of contexts. In particular, he can account for the universal import indefinites are claimed to get when they occur in a 'donkey-context'. However, there is a problem here. As Richards (1984) points out, if one accepts that *a donkey* in (9) and (10) has universal import, then surely it also has universal import in (11).

(11) Most men who own a donkey beat it.

However, if Kamp wants to account for this example using the same machinery, then he would run into problems. If the interpretation of *most* is to be parallel to that of *every*, then a quantified sentence with *most* as its highest quantifier is true if and only if most verifications of the first Discourse Representation constructed in response to sentence (11) can be extended to verifications for the

second Discourse Representation. However, this would not give *a donkey* in (11) full universal import. Rather, (11) would mean that most donkey-owning men beat most of the donkeys that they own.

There are a number of reasons for doubting the validity of Kamp's theory as a psychological theory. I will first give some arguments against the truth conditional approach in general. Soames (1984,163) claims that psychology and truth conditional semantics have nothing to do with each other. He writes:-

> Psychological theories are theories of the states and processes mediating sensory inputs and behavioral outputs. For mentalistic theories, the important states and processes are those occurring in the head (as opposed to the environment). ...
>
> Facts about truth conditions are not of this kind. To give the truth conditions of sentences is to specify the non-linguistic conditions that would make them true. ... [A] complete specification of the non-linguistic conditions under which they are true will not follow from a specification of mental states and processes, or a descriptions of the relationship between sensory input and behavioral output. Consequently, claims about the truth conditions of sentences are not (purely) psychological and linguistic semantics must be distinguished from theories of the mental states and processes underlying semantic competence.

However, apart from these higher level theoretical criticisms one can also give some more specific psychological arguments against Kamp's theory. All these arguments can also be made in some form against Heim's system. If we first look at the requirement posed by the claim that natural language is processed on a left-to-right basis, we see that the order in which the rules for the construction of Discourse Representation Structures are to be applied is determined by the syntactic analysis of the sentence. Thus, the component which contains these rules received full syntactic analyses as its input. Processing can therefore not take place on a left-to-right basis. This also applies to Heim's system who constructs her files on the basis of the logical form of a sentence. Second, Kamp's Discourse Representation Structures contain only information derived from the text, and do not also include information derived from background knowledge. Moreover, it is not at all clear that he could extend his theory to make the inclusion of background knowledge possible. Heim's theory has the edge

over Kamp's in this respect. She allows her files to keep track of objects which are contextually salient. She thus can introduce objects into the file which have not been explicitly introduced into the discourse.

4.2.5. Conclusion about the model-theoretic approach

Summarizing then, the model theoretic approach to the mental representation of discourse suffers from a number of essential shortcomings as a psychological theory. Since at least some of the criticisms I propose are not dependent on the specific theories developed thus far, I conclude that the model-theoretic approach in general has to be rejected as a fruitful approach to the problem of the mental representation of discourse.

4.3. The 'description' approach

The 'description' approach to the representation of discourse is exemplified by Webber (1979) and Richards (1984). The approach is essentially syntactic in that it heavily relies on the syntactic analysis of the sentences in a discourse. Both Webber and Richards give rules which can be used to construct certain descriptions based on the syntactic analyses of a given sentence. Webber calls them *invoking descriptions* whereas Richards uses the term *reconstruction terms*. Both use the constructed descriptions in an account of the interpretation of pronouns.

4.3.1. Webber's approach

Webber (1979;1-21 - 1-25) writes that an objective of discourse is the communication of a model from speaker to hearer. A discourse model is a set of entities naturally evoked by a discourse, the set of discourse entities, linked together by the relationships they participate in. Discourse entities are seen as hooks for descriptions,

the underlying conception being very much like that of KRL (Bobrow and Winograd, 1977; cf chapter 5). Discourse entities can be invoked in the hearer's discourse model either linguistically, perceptually, or inferentially. The first description to be attached to a discourse entity, i.e. the description which was attached when the discourse entity was invoked, is called the invoking description. What Webber sets out to do, is to identify those aspects of a text which are essential to forming appropriate invoking descriptions of discourse entities evoked by a text. Thus, one can regard Webber's invoking descriptions as descriptions of the mental objects which are constructed by the hearer in his mental representation of the discourse.

Webber gives a number of rules for the construction of discourse entities and their invoking descriptions. However, I will consider only one, as the point of the exercise is not to criticise the detailed theory Webber proposes, but rather to review the general approach of which Webber is one representative. It must be stressed that Webber considers a remarkably wide range of examples and different types of anaphora but in order to illustrate her approach, I will only consider her treatment of singular pronouns. Consider (12).

(12) Sam saw a cat.

Webber's rules take as input a logical translation of the sentence under consideration. The logical translation of (12) is (13).

(13) (Ex:cat).saw(Sam,x)

If we call sentence (12) S, then the rule which Webber gives takes the logical formula (13) as input and returns as output a discourse entity e1 together with the invoking description (14).

(14) the x: cat(x) & saw(Sam,x) & evoke(S,x)

(14) is informally rendered as the cat which Sam saw and which was mentioned in sentence S. The discourse entity e1 together with the invoking description can now be used as the interpretation of a subsequent pronoun.

Webber does not raise the issue about the relationship between evoked discourse entities and entities in the long-term knowledge base. However, given the fact that her work is firmly placed in the context of Artificial Intelligence, it is not too difficult to see how the frame-work she provides could be extended to do so. A first approximation might be to see the descriptions as linked to concepts in the long-term knowledge base, and discourse entities which correspond to long-term mental objects as pointing to their long-term counterparts.

Webber does not consider the question of how a given pronoun finds its antecedent. All she sets out to do, is to make clear how discourse entities are evoked, which at a later stage of the discourse can be used in the interpretation of a discourse. As a consequence, she also does not consider the question how a given discourse entity which is pronominalizable at one stage of the discourse gradually looses this status.

There is also a problem about the possibility of left-to-right processing of a sentence. Given that the input to the algorithm which constructs evoking descriptions is a logical translation of the sentence, this question of left-to-right processing reduces to the question of whether a logical translation of a sentence can be found in a left-to-right fashion. Given the fact that quantified NPs often appear in the verb phrase of a sentence, yet have to end up in the front of the logical translation of the sentence, it is not very likely that a sentence can be translated into a logical formula on a left-to-right basis.

Summarising then, Webber considers an impressive range of examples. Her theory is very precise and given that she works within the framework of AI, it seems straightforward to extend her theory to make inclusion of background information in the discourse models possible. It is less clear how left-to-right processing and the fact that discourse entities can loose their "pronominalizable" status can be accounted for.

4.3.2. Richards' approach

Another example of the description approach is Richards (1984). Richards's rules take as input not a logical translation of the sentence, but rather a syntactic analysis which is constructed along the lines laid down in Montague (1973). Thus, in order to illustrate how Richards' rule for the construction of reconstruction terms works, we have to a consider a syntactic analysis of (12). There is a problem here because there are in principle two different syntactic analyses of (12). One can either first combine the verb *saw* and the indefinite *a cat* to give a verb phrase which can then be combined with *Sam,* as in (15), or one can first combine *Sam* and *saw x* to give something which can be combined with *a cat* to give a sentence, as in (16)

```
(15)        Sam saw a cat
              |           |
            Sam     x saw a cat
                      |       |
                    a cat   x saw y
                    |     |
                  a   cat
```

```
(16)        Sam saw a cat
              |           |
            a cat     Sam saw y
            |   |     |       |
            a   cat   Sam   x saw y
```

The difference is not trivial because one gets different reconstruction terms for the different syntactic analysis. The algorithm for constructing reconstruction

terms relies on the syntactic analysis of the sentence. If we have a pronoun with an indefinite NP as its antecedent, then we construct a term interpreting the pronoun by looking at the syntactic analysis of the sentence the indefinite occurs in. The reconstruction term is then constructed out of all the predicates within the scope of the indefinite article in the syntactic analysis tree. Thus, if we take (15) as the syntactic analysis then the reconstruction term constructed in response to a pronoun in a following sentence which has *a cat* as its antecedent would be 'y[cat(y) & saw(x,y)]', a reconstruction term containing a free variable and hence not interpretable without additional machinery. On the other hand, if one takes the (16) as the syntactic analysis, then the reconstruction term would be 'y[cat(y) & saw(Sam,y)]'. Richards does not give any principled reasons why one should prefer one analysis to the other. In the rest of this section, we will, whenever relevant, take the analysis which would give the results Richards desires.

If we look at the reconstruction terms Richards constructs from a purely logical point of view, then we see that they syntactically behave just like logical constants, i.e. they appear in the same places in formulas as logical constants do. However, semantically they behave differently from ordinary logical constants. A sentence of the form 'F(x[Cx])' with x[Cx] a reconstruction term, is true in a model M just in case all individuals which have the property C also have the property F. Thus, if sentence (12) was followed in a discourse by sentence (17), then under Richards' analysis, (17) would means something like (18).

(17) It was black.

(18) All cats Sam saw were black.

This analysis predicts some counterintuitive results. Let us assume that Sam did indeed see a black cat, but that, unbeknownst to the speaker, he also saw another

cat at the same time which was white. According to Richards' analysis of (17), the speaker has then uttered an untruth. After all, only one of the cats which Sam saw was black, while the other was white. Thus, (17), under the analysis given in (18) would be false. However, this is a counter-intuitive result. While one may say that the speaker did not utter the whole truth, or even that, if she knew that Sam saw two cats, she had been deliberately misleading, it seems too strong to say that she lied.

Given the importance of this point for the appraisal of Richards' theory, let me make the same point in a slightly different way. Consider the following somewhat stilted discourse.

(19) Sam saw a cat.
It was black.
He saw that it was fighting with a white cat.

According to Richards' analysis, discourse (19) is inconsistent. Given that it follows from the fact that Sam saw a black cat fighting with a white cat that Sam saw both a white and a black cat, the speaker of (19) would claim at the same time that all the cats Sam saw were black, and also that one of the cats Sam saw was white. Hence, the speaker of (19) would utter an inconsistent set of statements. However, discourse (19) is perfectly acceptable.

Richards admits that his analysis may seem somewhat counterintuitive, but he argues that the theory should be judged not only on the basis of one example, but on the basis of a variety of examples. If a theory can be shown to give the right results in the case of a number of examples which are problematic for other theories, then, so the argument goes, we might be prepared to overlook a counterintuitive result in another case. While I can in general agree with this view, the argument should be used with caution. A syntactic theory which gives a perfect explanation of the formation of tag questions in English, but would make the wrong predictions

about the order of subject, verb and object in the main clause would be regarded with suspicion.

There are two advantages Richards sees for his theory. The first concerns the interpretation of plural pronouns. Richards argues that they will have to be analysed as universal quantifiers and that his theory of reconstruction captures this intuition. Richards' theory does indeed get the right results for the distributive interpretation of plural pronouns; the interpretation of the collective reading, which cannot be captured by a universal quantification over individuals, is left open.

The second main argument Richards sees for his theory is the treatment of the donkey sentences. His treatment enables one to account for the alleged universal import of pronouns which have as their antecedents an indefinite NP in a "donkey-position". Unlike Kamp's treatment of these examples which did not give the desired result for those cases where the indefinite occurred in a relative clause governed by a non-standard quantifier, such as *most,* Richards' treatment gives all the results he requires.

Richards sees the main objective of his theory as providing a translation of sentences of English into an appropriate logical language. In a sense, Richards is giving an extension of Montague's treatment of English. His theory therefore suffers from all the shortcomings of the model-theoretic approach to the mental representation of discourse which were discussed in sections 2. and 4.2. of this chapter.

4.3.3. An overall appraisal of the 'description' approach

It is clear that the roles of the Webber's invoking descriptions and Richards' reconstruction terms are very different and that this has important repercussions for

the value of the theory as a theory about the mental representation of discourse. Richards uses the reconstruction terms in providing a translation algorithm from English into a logical language. He thus runs into all the problems associated with the truth conditional approach to language. Webber sees the invoking descriptions as attached to discourse entities, mental constructs which can act as the interpretation of pronouns. She thus avoids the problems of the truth conditional approach.

Given that Webber's theory meets most of the criteria of adequay mentioned in the previous chapter, her theory is superior to those developed in the model-theoretic approach. Her theory suffers from two weaknesses however:- left-to-right processing and the fact that discourse entities can loose their "pronominalizable" status cannot be accounted for.

4.4. The mental models approach

The last approach to the mental representation of discourse is the mental models approach. There is a growing number of proponents of this theory (e.g. Karttunen, 1976; Fauconnier, 1979; Seuren, 1985; Johnson-Laird, for example 1983). All agree that the aim of a lot of discourse is the transfer of information from speaker to hearer and that the processes which led the speaker to produce the discourse and the processes by which the hearer understands the discourse are of crucial importance to the study of human language. There is therefore an intimate connection between the mental models approach and procedural semantics paradigm discussed in chapter 2.

Most theories proposed in this framework lack the formal rigour of the other approaches to discourse representation. Moreover, there is also lack of agreement about the exact nature of the models. Because of this I will not discuss any of the formal proposals in detail as I did in the case of the model-theoretic and the

description approach. Rather, I will concentrate on some of the properties which have been ascribed to mental models.

An essential feature of the mental models approach which distinguishes it from other approaches to the mental representation of discourse lies in the fact that the mental representations which underlie the processes of comprehending and producing discourses, are regarded as having semantic status themselves and not in need of any further interpretation. Unlike the truth conditional approach where the discourse representations have to be embedded into standard logical models and derive their semantic status from this embedding function, the structures postulated in the mental models approach are themselves semantic entities. Seuren (1985) makes this point forcefully when he writes:-

> possible worlds should be done away with in semantics; their role should be taken over by a theory of mental machinery accounting for everything, and a great deal more, that possible worlds were meant to account for.

Thus, in Seuren's theory, and in this respect he is representative of the mental models approach, mental models take over the role of the possible worlds in standard formal semantics. Consequently, in the mental models approach once the task of establishing a relationship between discourses and mental models has been fulfilled, one has given a complete semantic theory.

Johnson-Laird (1984, chapter 15) lists a number of properties of mental models. He gives six principles concerning the nature of mental models and three principles concerning the concepts which mental models can embody. The latter need not concern us here. The six principles about the nature of mental models are:-

1. the principle of computability,
   mental models, and the machinery for constructing and interpreting them, are computable

2. the principle of finitism,
   a mental model must be finite in size and cannot directly represent an infinite domain

3. the principle of constructivism,
   a mental model is constructed from tokens arranged in a particular structure to represent a state of affairs.

4. the principle of economy in models,
   a description of a single state of affairs is represented by a single mental model even if the description is incomplete or indeterminate

5. mental models can directly represent indeterminacies if and only if their use is not computationally intractable, i.e. there is not a exponential explosion in complexity.

6. the principle of structural identity,
   the structures of mental models are identical to the structures of the states of affairs, whether perceived or conceived, that the models represent.

The first three principles are more or less equivalent to the psychological criteria of adequacy for representational systems for the mental representation of discourse; the structures postulated to represent discourse have to be psychologically plausible. The fourth principle is equally plausible and therefore I will not discuss it in any detail. Indeed, the findings of Stenning (1981) that people tend to make specific assumptions even if they are not warranted to do so by the actual discourse, which were mentioned in section 3.2. of this chapter, provide experimental support for this principle. Principle 5. is almost a corollary of the principles 1., 2. and 4., and is therefore also plausible.

The principle of structural identity is central to Johnson-Laird's conception of mental models. Johnson-Laird (1983,419-22) uses it to clarify the distinction between the mental models he postulates and the partitioned semantic networks of Hendrix (1975,1979). He dismisses the semantic network approach partly because it does not obey the principle of structural identity. There is no structural analogy between an

(external) state of affairs in which an assertion would be true and the semantic network constructed on the basis of the assertion. There has to be a structural identity between structures used in ontology and those postulated in epistemology and semantic networks do not have this.

In section 2.1. of chapter 2 I put forward an evolutionary argument for the assumption that the representations which humans construct of reality are likely to bear a close resemblance to reality itself. I thus put forward an evolutionary argument for the principle of structural identity. However, it seems to me that the practical usefulness of the principle is somewhat restricted. The claim that the structure of a mental model is identical to the structure of the state of affairs it represents is only a substantial claim if the structure of states of affairs can be clearly specified independently of our cognitive apparatus. But there is no such specification. Indeed, ontologists have argued over the centuries about the structure of reality, and although the debate might be slightly less fashionable these days, the problem has never been solved, as far as I am aware. One may actually doubt that the question can ever be solved. Given Locke's ´veil of perception' (section 3.1.1. of chapter 2), it is impossible to get direct access to the world. Johnson-Laird seems to agree when he writes:-

> In short, our view of the world is causally dependent both on the way the world is and on the way we are. There is an obvious but important corollary:- all our knowledge of the world depends on our ability to construct models of it.
>
> Johnson-Laird (1983;402)

Thus, we have no independent means of determining the structure of states of affairs independently of ourselves. The principle of structural identity, true as it may be, is therefore of little practical use when we are trying to determine the structure of our mental representations.

There are two ways forward. One can either postulate what the structure of reality is, or one can make proposals about the structure of the mental models. One could argue that logicians take the first option when they construct a model theoretic interpretation of the language they are concerned with. Thus, when one is dealing with a non-modal language, all one needs in one's ontology is individuals and truth-values [9]. All the other elements one needs for interpretating the language, such as the interpretations of predicates, can be built up out of the set of individuals by standard set-theoretical means. When one is dealing with a modal language, then one needs a set of possible worlds, or something similar, to build up all the constructs one needs for interpreting the language [10].

There are however two problems with this option. First, logicians do not agree among themselves about the proper structure of their models. Quine's crusade against the use of possible worlds in the semantics for modal logics has already been mentioned in footnote 1, and more recently Barwise and Perry (1983) have proposed an ontology which is completely different from the one normally used by logicians. Secondly, unlike logicians, cognitive scientists cannot just postulate the structure of reality in order to determine the structure of mental models. They are restricted in their options by certain empirical considerations, some of which are for example embodied in the other principles Johnson-Laird proposes.

Given the failure of the first option, one is forced to take the second option. rather than postulating what the structure of reality is, one postulates the structure of the mental models. This problem has of course been extensively studied in Artificial Intelligence under the name of knowledge representation, and various knowledge representation schemas have been proposed (cf Barr and Feigenbaum 1982; chapter III). Although we also not have direct access to our own mental

representations, and although we therefore cannot directly determine the structure of our mental representations, we have the advantage that we can experiment with different proposals by implementing them on a computer. In chapter 5., I will return to the question of knowledge representation and propose a knowledge representation schema which is very similar to Minsky's frames (Minsky, 1975), and Bobrow and Winograd's Knowledge Representation Language (Bobrow and Winograd, 1977).

In the previous discussion, we have drifted from the question of the structure of the mental representation of discourse to the question of mental representations in general. The question naturally arises what the relationship between the two is, and it is here that the fifth axiom of procedural semantics, mentioned in section 2 of chapter 2, becomes relevant:- the most appropriate formalisms for building theories of language are those that deal explicitly with the structure of knowledge and the processes using it. The principle can be argued for as follows - language is used as a vehicle for the transfer of information from speaker to hearer. Speakers have information available in their long-term knowledge stores which they want to transfer to their hearers. Hearers, on the other hand, receive information in a discourse which, if we restrict ourselves to the ideal situation and abstract from factors such as memory limitations, will become part of their long-term knowledge bases. If we apply Occam's razor, then it can be expected that the representations used in discourse production and the representations constructed in discourse comprehension have the same structure as the representations used to store long-term knowledge. It thus follows that the question about discourse representation reduces to the problem of knowledge representation, a question to which I will return in chapter 5.

4.5. Conclusion

The first task facing those interested in the mental representation of discourse is to specify the structure of the long-term knowledge base in the human language processor, since it can be expected that the mental representations used in discourse production and comprehension are of the same form as those in the long-term knowledge base. However, having done so, they will also have to specify the relationship between certain expressions in the language and parts of the structures in the knowledge base. After all, if language is a medium for transferring information from speaker to hearer, and the representations constructed in response to language reflect the structure of the knowledge base, then there have to be systematic correlations between certain expressions in the language and certain structures in the knowledge base. The first problem will be discussed in chapter 5., the second in chapter 6.

5. Summary and conclusion

In this chapter I discussed two theories about the overall architecture of the human language processor. In the next chapter I will present another model of the human language processor which resembles Sanford and Garrod's model. My model however avoids some of the problems raised in connection with their model.

I also discussed some theories about the specific representations used in discourse production and comprehension. I came to the conclusion that it can be expected that the representations in question will be of the same structure as the representations in the long-term knowledge base. I will therefore introduce a knowledge representation schema in chapter 5., while in chapter 6. I will discuss certain linguistic expressions to see how their behaviour in discourse could be

accounted for using the representation schema developed in chapter 5.

Footnotes:-

1. The model theoretic interpretation of intensional logic is unproblematic only from a logical point of view. Indeed, Quine has spent a good deal of his career arguing that the model-theoretic interpretation of modal logics as developed by Kripke is not without its problems from an ontological point of view as it relies on the notion of a possible world. Quine argues that a commitment to this notion leads to undesirable extensions of the ontology.

2. A number of Montague's intellectual heirs (e.g. Partee, Dowty, Kaplan) have attempted to extend the basic Montague programme to deal with world-knowledge. However, in their work world-knowledge only plays a role in the interpretation of propositions and not in the construction of propositions from linguistic input. Therefore, most of these arguments also hold against their position, as it now stands. It is an open question whether their efforts can be extended to address these problems as well.

3. Sanford and Garrod distinguish between Explicit and Implicit Focus because pronouns, they claim, can only be used to trigger searches of explicit focus. It has to be noted that this way of drawing the distinction is incorrect. Yule (1982) reported a number of examples of pronouns which refer to entities which have not been explicitly introduced in the discourse but have to be derived from activated background knowledge. I will return to Yule's examples and pronominalization in general in chapter 6. For the time being I will restrict myself to definite descriptions and the memory searches they trigger according to Sanford and Garrod.

4. A syllogism is an argument with two premises, each of which has one of the four following forms.

All x are y.
Some x are y.
No x are y.
Some x are not y.

One of the terms (i.e. the x or y in the above formulas), the so-called middle term, is the same in both premises. The conclusion which also takes one of the four forms, then contains the other two terms in the premises. An example syllogism is the following:-

Some x are y
All y are z

Some x are z

5. Mark Steedman pointed out to me that this is not a necessary consequence of the overall approach. In the case of syllogisms, one can imagine that the models which change the initial model do so in such a way as to leave the premises true on the model. Thus, one can imagine procedures which change the model in such a way that the premises remain true. However, it seems to me that it is harder to see how one could do so in, for example, spatial reasoning.

6. Johnson-Laird qualifies his position in a number of places. The point of the procedures is to show *in principle* how one model can 'stand for' many models. However, Johnson-Laird admits that in practice the extent to which the procedures can be carried out is restricted by for example the limits on working memory

(Johnson-Laird, 1983; 264). He also point to the Gricean principles which speakers and hearers obey. As a consequence, radical revision of a model is not often called for in ordinary discourse. (Johnson-Laird, 1983; 164).

7. Sentence (8) feels somewhat strange and is much improved if one expands the verb phrase, as in (1') and (1").

(1') A miner who owns a donkey will beat it.

(1") A miner who owns a donkey has the right to beat it.

8. The actual truth definition Kamp gives is rather complex. It contains basically all the machinery of a standard Tarskian truth definition with separate clauses for atomic formulas and the different connectives. However, as van Benthem and van Eijck (1982) point out, this complexity is necessary. The simplest definition of extentability would be:- discourse D is true in model M if and only if Discourse Representation Structure DRS can be extended to M. But this is not adequate as one can prove that it can only apply to purely existential sentences and would not give the right results for universal sentences.

9. In principle one could do away with truth values. Tarski's original definition of an interpretation for predicate calculus did not make use of truth values. The idea is to see the interpretation of a formula as the set of those assignments of individuals to variables which satisfy the formula. A formula can then be defined to be true just in case its interpretation is the set of all assignments, and false if its interpretation is empty.

10. In fact, it is to also construct possible worlds from the set of possible individuals and truth values by standard set theoretical means (Cf Reichgelt, 1981). The basic idea is to treat a possible world as a standard interpretation for a non-modal first order language.

Chapter 4:- The architecture of the human language processor

1. Introduction

In this chapter I will develop a psychologically motivated computational model for the human language processor. Earlier versions of the model can be found in Reichgelt (1982) and Shadbolt (1983). The model is intended to model the more 'semantic' and 'pragmatic' aspects of the processor, and is not intended as a theory about how humans parse sentences. The question about the specific representations which people use or construct in discourse understanding will be largely left until chapter 5.

1.1. Knowledge activation

Using language in discourse involves the activation of knowledge. This is obvious from the speaker's point of view. After all, a speaker wants to transfer certain information which is stored in her long-term knowledge base to the hearer. In order to be able to do so, this information will have to be made available for the language generation processes, i.e. it will have to be stored in a more activated format. Moreover, since speakers generally cannot transfer the information in one utterance, but have to transfer it in different utterances, they have to be able to keep track of what information has already been transferred. So, the information has to remain available in its totality during the entire discourse in a more accessible, i.e. more activated, format.

Understanding language also involves the activation of knowledge. Comprehending a discourse often is impossible without an understanding of the background against which the discourse is set. Various authors have noted that the

early stages of discourse often have the function of making the hearer realize against which background the discourse is to be understood. (Cf. Garrod & Sanford, 1982). The early stages of discourse often have no other role than enabling the hearer to activate the appropriate background knowledge. Another reason for assuming that understanding a discourse requires the activation of knowledge emerges from a closer look at what a hearer has to do during discourse comprehension. The point of the discourse is the transfer of information from speaker to hearer. The hearer is therefore faced with the task of reconstructing in his own mind the information which led the speaker to produce the discourse. This means that hearers not only have to understand each of the single utterances speakers produce; they also have to integrate the interpretations of the different utterances into a consistent and coherent informational structure. Clearly, this presupposes that the hearer keeps information provided earlier in the discourse available in a format which makes integration of new information easy, i.e. in an activated format.

The model I will present here distinguishes between three functionally different components for modelling knowledge and its activation. The first component is the discourse model. It contains a representation of the interpretation of the ongoing discourse. A second component, the general epistemic model, is the representation of the long-term knowledge processors bring to a discourse. A third component, the discourse specific epistemic model, contains knowledge activated on the basis of previous discourse as well as the long-term background knowledge which is necessary for the production or comprehension of the discourse. The distinction between the three components has to be seen as a distinction between various degrees of activation of the different pieces of information or knowledge. The information in the discourse model can be thought of as more highly activated than the information in the discourse specific epistemic model and similarly for the discourse specific

epistemic model and the general epistemic model.

The distinctions are thus functional rather than ontological. There are various ways in which one can represent degrees of activation of knowledge in a theory of the human language processor. One way would be to have an index representing the degree of activation attached to each piece of information represented in the system modelling the processor. Another possibility is the one I adopted here. Rather than having a direct representation of the degree of activation, I represent the most activated information as being in the discourse model, the least activated information as being in the general epistemic model, and pieces of information which are activated to an intermediate degree as being in the discourse specific epistemic model. This gives an indirect method of stating the degree of activation of a particular piece of knowledge. The main drawback of this way is that it appears that there are only three possible degrees. Information is either in the discourse model, or in the discourse specific epistemic model, or the general epistemic model. This is an abstraction. There are probably many more degrees of activation. It is even likely that the scale of activation is continuous, thus giving an infinite number of degrees of activation.

1.2. Comparison with Sanford and Garrod's model

There is a large similarity between the model presented here and the model proposed by Sanford and Garrod which I discussed in section 3.1. of chapter 3. If one compares the various components Sanford and Garrod distinguish and the functional "components" in the model presented here, then one will notice a large similarity between the two. At least at first sight, there is a correspondence between Sanford and Garrod's Explicit Focus and the discourse model, between Long-Term Semantic Memory and the general epistemic model. The role of the discourse specific epistemic

model is taken in Sanford and Garrod's model in part by Implicit Focus and in part by Long-Term Text Memory. I argued in section 3.1. of chapter 3 that Sanford and Garrod's distinction between Implicit Focus and Long-Term Text Memory could not be maintained. Since there is one component in my model which combines the roles of the different components in Sanford and Garrod's model, my model is not susceptible to this criticism. I will return to this point in section 3.

There is also a major difference between Sanford and Garrod's views and the ones presented here concerning the status of the different components in the model. Sanford and Garrod see the different components as memory partitions and they define a memory partition (1981;158):-

> as independently addressable and capable of being treated by the processor as a distinct search domain.

I see the various components in my model as ordered along an axis of activation. I thus make a quantitative difference between the various components, whereas Sanford and Garrod's distinction can be seen as qualitative. It has to be pointed out however that Sanford and Garrod certainly seem to be a bit ambivalent in their position on the status of the different components in the model. While the above definition of a memory parition suggests a rather "monolithic" view of the various components, their discussion of Explicit Focus certainly implies that not all information in the various components has the same status, but that at least within Explicit Focus there are various degrees of activation. After all, entities in Explicit Focus are seen as taking up computational space, and the more space they take up, the more foregrounded, or in my terms activated, they are. Sanford and Garrod also say that Implicit Focus is simply an activated part of Long-Term Semantic Memory. In other words, the notion of activation of knowledge which is central to my account creeps into Sanford and Garrod's model in a number of places.

Although there thus are some differences between Sanford and Garrod's model and the present one, these differences do not yield any empirically testable consequences:- there are no psychological experiments which would allow one to distinguish the two theories. The only reason for preferring the present theory is the principle of parsimony. Since the present theory postulates fewer components than Sanford and Garrod's theory, it has to be preferred. Notice incidentally that the absence of empirically testable differences implies that the wealth of empirical evidence which Sanford and Garrod put forward in defence of their model also applies to mine.

1.3. Speakers and hearers

The model postulates the same components for speakers and hearers. This may look counter-intuitive especially since from an intuitive point of view speaking and hearing are two different activities. However, the model allows one to explain the difference between speaking and understanding in terms of the interaction or flow of information between the different components. Although the (static) components are the same in speaking and hearing, the (dynamic) processes are different.

One reason for preferring this solution to one which postulates different models for speaker and hearer is Occam's razor. It is more parsimonious to assume that the same components can model both speaker and hearer. However, it is possible that this assumption has to be given up later in face of the facts.

There is another argument for assuming that the components have to be very similar for speaker and hearer. In a lot of discourses discourse participants switch roles quite regularly. Speakers become hearers and hearers become speakers. This suggests that the structure of the human language understander and the structure of

the human language generator have are very similar. By having the same components for the speaker and the hearer, but by changing the processes underlying speaking and hearing, the above model is in accordance with the ease with which discourse participants change roles.

2. The general epistemic model

The first component in the model for the human language processor is the *general epistemic model.* It can be identified with our long-term knowledge store. Sanford and Garrod (1981) use the term long-term semantic memory. The general epistemic model contains the long-term knowledge which speakers and hearers bring to a discourse.

Since the other components of the system are to be seen as modelling further degrees of activation of knowledge, the remarks I make in this section about the structure of the general epistemic model also apply to the other components. In this section I will sketch only in the roughest outline the way in which knowledge is represented in the processor. I will return to this question in the next chapter.

2.1. Conceptual entities

In chapter 5. I will follow Bobrow and Winograd (1977) and opt for an organization of the knowledge base around conceptual entities rather than the structuring of knowledge around sets of facts, each referring to one or more objects. The best way to see the basic unit in the knowledge base is as a focal point plus a set of properties or beliefs hanging off it. In a similar vein, Collins and Quillian (1972) write :-

> In fact, human concepts are probably more like hooks or nodes in a network from which many different properties hang.

In order to avoid confusion, I will use *conceptual base* for the focal point and

the term *conceptual entity* for the whole basic unit, i.e. the conceptual base plus the set of beliefs.

It is important to stress here that the notion of conceptual entity which I propose is a relatively rich one. A conceptual entity is not just the mental counterpart of an object in the world, as used by logicians when they define model-theories for logics. The notion of a conceptual entity is much richer: a conceptual entity is a relatively large chunk of knowledge. I will clarify the notion of conceptual entity in chapter 5.

2.2. Embedded models

In chapter 6. I will argue that whenever a speaker introduces a new object into the discourse with the intention of providing her hearer with new information about it, the choice of the linguistic expression partly depends on the beliefs about that object she ascribes to her hearer. It follows that part of the knowledge in conceptual entities has to be the processor's beliefs about the beliefs of others. There thus must be various sets of beliefs in conceptual entities. The set of beliefs in a conceptual entity to which the processor itself ascribes will be called the *primary* model or perspective. The set of beliefs which the processor thinks some other processor x ascribes to is the *secondary* model or perspective for x. The tertiary model then is the set of beliefs in an conceptual entity which represent the beliefs which the processor thinks some other processor x believes that the processor itself ascribes to the conceptual entity, etc.

It is useful to reiterate that the general epistemic model models the least degree of activation a piece of knowledge can have. Both the discourse specific epistemic model and the discourse model model higher degrees of knowledge

activation. They have the same structure as the general epistemic model, however. This of course applies to the existence of embedded models as well. So, there are also embedded models representing the beliefs of the processor itself and the beliefs it ascribes to others etc. in the discourse specific epistemic model and the discourse model.

The idea of embedded models put forward here is a generalization of an idea proposed by various authors, such as Fauconnier (1979), Johnson-Laird (1983;430-438), Shadbolt (1983), and Seuren (1985). Hendrix' work on partioned networks can be interpreted as introducing similar machinery. (Hendrix, 1975 1979). Fauconnier did the most detailed work on embedded models. He does not explicitly use them to represent the beliefs one processor ascribes to another however; rather, he uses them to account for a number of expressions including the verbs of propositional attitude, negation, conditionals etc. Roughly, the idea is that certain expressions in natural language are "space-creating" expressions:- in response to them the hearer will create an embedded model. If we restrict ourselves to the verbs of propositional attitude, then the embedded models constructed in response to them are representations of the beliefs, wishes, doubts, hopes, or any other propositional attitude that the object referred to in the (grammatical) subject of the verb of propositional attitude may have. The various components in my model always contain embedded models representing the beliefs processors ascribe to their fellow interlocutors. In this respect, my ideas are a generalization. In chapter 6. I will make use of the idea of embedded models.

2.3. Mutual knowledge

The notion of mutual knowledge [1] will also be shown to be of crucial importance to an explanation of discourse production and comprehension in chapter 6. One can use

the above idea of embedded models to define a notion of mutual knowledge:- mutual knowledge about an object can be read off from the different perspectives by looking at what is constant between the different levels. So, what a processor thinks is mutually known is what is constant between all levels. What a processor thinks somebody else thinks is mutually known is what is constant from level 2 downwards.

There is of course a problem with the recursion in the definition of the various levels which are necessary in the account of mutual knowledge outlined above. Obviously, there has to be an end to the number of levels. If there was not, our beliefs would be infinite sets and this can hardly be maintained as a psychological possibility. Clark and Marshall (1981) speak in this context of the mutual knowledge paradox. Mutual knowledge seems to involve knowledge of an infinite number of conditions. On the other hand, definite reference can only be explained in terms of mutual knowledge (cf chapter 6). People clearly can handle expressions which are used to refer definitely in a finite amount of time. But people cannot handle an infinite amount of information in a finite time. We are therefore forced to admit that mutual knowledge has to be expressible in a finite way.

In the above system mutual knowledge can be implemented in a finite knowledge base by stopping when there is no difference between the sets of beliefs on level n and those on level n + 1 except for the fact that level n contains level n + 1 as a substructure. In other words, when there is no difference between one level and the next, then there is no reason to go down any further in the representation. Lewis (1969) proposes a similar rule. However, the above definition should not be taken as a claim about how mutual knowledge is acquired and stored in the human knowledge base. I will follow Clark and Marshall in assuming that humans use heuristics to decide whether something is mutually known, and may not use the different

embedded models in the way I suggested here when it comes to determining mutual knowledge.

This account of mutual knowledge for particular conceptual entities is processor-dependent, as it should given the philosophical points made in chapter 2. What you think is mutually known may differ from what I think is mutually known or even from what I think you think is mutually known. Clark and Marshall (1981) call this account a "one-sided definition of mutual knowledge". Schiffer (1972;30-1) originally defined mutual knowledge in a processor-independent way [2]. But as Clark and Marshall rightly point out this definition represents mutual knowledge as an omniscient observer would see it, an observer who can look into people's heads and determine what they know. But obviously people are not omniscient. Language users can only determine for themselves what they think they and their fellow interlocutors mutually know. They can only make assumptions about the knowledge of others and have no direct access to it. It is for this reason that Clark and Marshall claim that the one-sided definition of mutual knowledge is more useful.

Clark and Marshall discuss some of the heuristics people might employ in determining whether something is mutually known and in this discussion the thesis of cognitive similarity plays an important role. They claim that mutual knowledge can be seen as a function of a number of factors. The main factor is a *mutual knowledge induction schema,* which is an adaptation of a similar schema in Lewis (1969):-

a and b mutually know that p iff
some state of affairs g holds such that
1. a and b have reason to believe that g holds
2. g indicates to a and to b that
   each has reason to believe that g holds
3. g indicates to a and b that p.

The problem with this formulation of the schema is that it again is processor-independent. Although Clark and Marshall point to the greater usefulness of one-sided definition of mutual knowledge, they are a little too careless in the formulation of the induction schema. They admit that they will loosely speak about mutual knowledge as such. The problem, however, is that it is not entirely trivial to reformulate the induction schema in a processor-dependent way. A possible formulation is:-

a believes that a and b mutually know that p
iff a thinks that some state of affairs g holds such that
1. a has reason to believe that g holds
   and has reason to believe that
   b has also reason to believe that g holds.
2. a believes that g indicates to both a and b
   that each has reason to believe that g holds.
3. g indicates to a that p and a
   believes that g indicates to b that p.

In both induction schemata, there are two variables which have to be specified in order for the induction to be possible. The first concerns the state of affairs g itself, or the *basis*, as Clark and Marshall call it. (g stands for ground). The second variable concerns the reasons a has for believing that b has reason to believe that g holds and moreover to believe that a believes that g holds. This variable is called '(auxiliary) assumptions'. Clearly, the thesis of cognitive similarity introduced in chapter 2. is a very elementary auxiliary assumption.

There is a certain trade-off between the basis and the auxiliary assumptions. The stronger the basis, the less auxiliary assumptions I will have to make, and vice versa, the more auxiliary assumptions I make, the less strong the basis I can rely on. So, if we are looking at each other across the table with a candle between us on the table, then the physical copresence of the candle is obvious and the auxiliary assumptions I have to make are relatively weak. If the candle is behind you, however,

and I saw that you saw it before, the physical copresence is less clear. In order to suppose that the candle is mutually known, I not only have to assume that you actually noticed it when you looked at it, I also have to assume that you remember seeing it.

3. The discourse specific epistemic model

Although understanding a discourse usually requires a lot of background knowledge, I showed in chapter 2. that at a particular stage of discourse only a small part of the general epistemic model is activated. Understanding a particular discourse requires access to only a limited part of the long-term knowledge base. The discourse specific epistemic model is intended to model this.

The discourse specific epistemic model has more roles however. It will be clear that it is not always the case that we remember all the information we receive in a discourse in the long term. We tend to forget quite a lot of what we have been told. But, while we are participating in a discourse, we generally have an accurate idea of what was said in the preceding discourse, independently of whether it will eventually be integrated in our long-term knowledge store. Of course, when we participate in a long discourse, we do not retain all the information we have received. Our working memory is simply too small. In general, however, we will have a fairly accurate idea of what the speaker is talking about at the moment. How much information and what type of information one retains depends on a number of factors which I will not go into. For the moment, it is sufficient to note that we have a discourse memory where we retain information relatively independently of whether it is ever going to be integrated in the general epistemic model. Part of the function of the discourse specific epistemic model is to act as a store for information received earlier in a discourse. So, the discourse specific epistemic model receives its

information from at least two sources. Some of it comes from the general epistemic model and is the activated background knowledge. Other information comes from previous discourse.

The above hypothesis makes some predictions about the way texts are stored in human memory. If the discourse specific epistemic model contains information derived from the text itself and activated relevant background information, then memory for texts should show this. Recall for texts should show that people produce both information from the text itself and information which has been retrieved out of the general epistemic model. The psychological data discussed in section 3.2 of chapter 2. confirm this prediction.

3.1. A dilemma

The discourse specific epistemic model combines the roles of implicit focus and long-term text memory in Sanford and Garrod's model. Although neither psychological data on text memory and text understanding nor the behaviour of definite descriptions warrant the distinction, Sanford and Garrod made the point that it was important for the hearer to distinguish between information he received in the discourse and information he retrieved from his long-term knowledge store. As I will argue below, this is certainly true for the speaker, and since in a lot of discourses hearers can become speakers, the distinction is also important for hearers. In the rest of this section I will show how one can make this distinction for speakers. The arguments however directly apply to hearers as well.

That speakers have to distinguish between information they have already uttered in the discourse and information which came from their general epistemic models, can be shown by looking at discourses in which more than one object has

been introduced. Suppose the speaker wants to provide her hearer with information about one of the objects. Suppose moreover that she cannot use a pronoun to do so, because none of the objects has been sufficiently foregrounded. Suppose that she also cannot use a proper name because she does not believe that her hearer knows the intended object under this name. In that case, the speaker will have to use a definite description constructed on the basis of information which she can suppose her hearer to know. Now, she can assume that her hearer has available information which was uttered eralier in the discourse, whereas she cannot (generally) make this assumption for information which comes from here general epistemic model and has not yet been uttered [3]. Thus, she can use information uttered earlier in the discourse to construct a definite description but not information which comes from her general epistemic model. As an illustration, consider discourse (1).

(1) Yesterday, I met an elderly woman
and her daughter in Waverley station.
The daughter was wearing a pink dress.

A speaker cannot in general use the definite description *the woman in the pink dress* or the pronoun *she* immediately after the first sentence to refer to one of the objects introduced into the discourse. The speaker will only refer successfully if she uses a 'disambiguating' description. But she can only construct such a description if she distinguishes the information in the discourse, which is also available for the hearer, from the information which came into the discourse specific epistemic model from the general epistemic model, and which they cannot suppose their hearer to also have. We now face a dilemma:- on the one hand, I have rejected the idea of having two components, one for storing information derived from the text, another for storing activated background information, as proposed by Sanford and Garrod. On the other hand, it seems to be necessary to be able to distinguish between the two types of information. In the next subsection I will use the idea of embedded models to solve

this dilemma.

3.2. Embedded models again

The dilemma can be solved by using the idea of embedded models introduced in section 2.2. There I used this idea to account for mutual knowledge. Since the discourse specific epistemic model has the same structure as the general epistemic model, it will also have embedded models and we can use these to account for activated mutual knowledge. Under ideal circumstances, speakers will assume that the information which they have uttered in the discourse has been transferred to the hearer. Given no indication to the contrary, the speaker will assume that the hearer believes what she has said. So, after an utterance of a sentence the speaker will suppose that the hearer now believes the information in the sentence. In the terminology developed above, the information will be part of the speaker's secondary discourse specific epistemic model.

However, speakers will under ideal circumstances also believe that hearers believe them to hold the views which have been expressed in the sentence. That is, the speaker will assume that the hearer now realizes that the speaker believes what has just been said. In other words, the tertiary model in the speaker's discourse specific epistemic model will be similar to the secondary model. This of course makes it mutually known according to the speaker. Obviously, something similar happens for the hearer.

The distinction between information which comes from the speaker's general epistemic model and has not yet been verbalized, and the information which has been verbalized in the discourse can now be rephrased in terms of mutual knowledge. The information coming from the speaker's general epistemic model exists in her primary

discourse specific epistemic model alone and is not yet mutually known and can therefore not be used in constructing the definite description in discourse (1). The information which has been uttered in the discourse, on the other hand, can be assumed to be mutually known and can therefore be used in discourse (1). The conclusion is that by using embedded models, which we need for independent reasons anyway, we can distinguish between information which the hearer received in the discourse and information which comes from his long-term knowledge base alone. The first type of information is mutually known in the discourse specific epistemic model, whereas the second type is not. We thus see that the distinction which led Sanford and Garrod to distinguish between implicit focus and long-term text memory can be made in one component using machinery which was independently motivated.

4. The discourse model

The last component in the system for representing the human language processor is the discourse model. It models yet a further degree of activation and contains the most activated knowledge relevant to the discourse under consideration. Its main use is in the explanation of pronouns, and it will used thus in chapter 6.

For the speaker the discourse model contains a representation of the informational content of the utterance she is currently producing. For the hearer, the discourse model contains a representation of the informational content of the utterance he is currently processing. The exact moments a representation of the content is placed in the discourse model are slightly different for speaker and hearer. A speaker knows in a sense what the informational content of her utterance will be before she actually produces it. A speaker plans and knows what she is going to say in advance. A hearer on the other hand is in a less priviliged position. because of factors such as his knowledge about the speaker, his beliefs about the speaker's

beliefs about him, his knowledge about the topic under discussion etc, the hearer may have a fairly accurate idea about what the speaker is going to say, but he cannot be sure until the speaker has actually produced the utterance.

It follows from the above considerations that for the speaker material is transferred to the discourse model before it is verbalized in the discourse. Verbalization thus is a process which takes as input material in the speaker's discourse model and produces as output a utterance. Obviously, there has to be some feedback mechanism. While the utterance is produced, speakers will keep track of what they are saying, simply because they will have to check whether the utterances produced are actually what they want to say. It has been suggested to me that this indicates that a speaker may not put material in her discourse model until after the utterance has been produced. However, given the fact that the speaker sometimes does not notice that what she actually says is not what she wanted to say, I would suggest that what a speaker does when she is producing an utterance is compare an interpretation of the utterance she is producing with material already available in her discourse model and thus checks the correctness of the actual utterance.

Two things have to be noted. First, I do not want to imply that the speaker places all information which she wants to transfer in one particular discourse, or even all information she want to transfer in one utterance, in the discourse model at once. Verbalization probably takes place on a left to right basis and one reason may be the way in which the information is activated. Consider for instance sentence (2), occurring as the first sentence of a discourse.

(2) You remember this guy we met in the pub
just off Princesstreet last Friday +
or was it Thursday, I can't remember +
anyway, that guy found himself a job today.

The speaker knows in advance the chunk of information she wants to transmit to her hearer. However, she has to use language and thus is forced to encode it in a linguistic string. The problem is that in general she will not be able to transfer this chunk of information in one sentence. She therefore has to package the information. The first step the speaker has to take in the generation of this utterance, is to set up a new discourse object in her discourse model. She has to activate into her discourse model a token corresponding to the mental object which she believes to be a mental representation of the object she wants to provide information about. She then lexicalizes this and thus instructs her hearer to set up a corresponding discourse object, i.e. the speaker instructs the hearer to search his knowledge base for a corresponding mental object which he is then to activate as well. Presumably, one of the reasons for using the complex description in (2) is to make the hearer's search of his general epistemic model easier. The speaker then attaches the 'property' of finding a job to this discourse object and lexicalizes this bit of information, and instructs her hearer to do so as well. If everything goes all right, then the hearer will also attache the property 'finding a job' to his discourse object. So, the fact the speaker has to use language to transmit information and therefore cannot transmit the complete chunk of information at exactly one moment in time (as she presumably could do if she could communicate telepathically), means that she has to package it. This is reflected in the way the information is activated into the discourse model.

The second thing to note is that the speaker also has to keep a record of what she has said and what she can as a consequence assume to be activated in the hearer's mind. As will be argued in chapter 6., the need for this is most clearly shown by the behaviour of pronouns. In the present system, this is modelled in the same way as the distinction between what which the speaker presumes to be mutually

known and what she believes to come exclusively from her own general epistemic model is modelled in the discourse specific epistemic model.

5. Dynamics of the model

Having discussed the various components in the model, it is now time to turn to the dynamics of the model. The discussion will be split in two parts. in the first part I will discuss how the processes underlying the production of discourse are modelled in the system, i.e. how the speaker is modelled, and in the second those underlying the comprehension of a piece of text, i.e. how the hearer is modelled.

5.1. Speakers

In transactional language use the speaker has the intention of transmitting some information to her hearer. The processes by which she does so, are slightly different depending on whether the utterance she is producing is the first utterance in a new discourse or a new episode in a discourse, or whether the utterance is part of an ongoing episode in discourse. I will discuss the former case first.

5.1.1. New episodes

In order to be able to transfer information which she supposes to be new to her hearer, the speaker will first have to activate this information. In a new (episode of a) discourse, before the speaker has produced an utterance, this information resides in the speaker's general epistemic model. Activation of information is modelled as a transfer from the general epistemic model into the discourse specific epistemic model. The speaker then has to package this information on the basis of what knowledge she ascribes to her hearer. The packaging is reflected in the order in which the information is activated to an even higher degree and placed into the discourse

model. From the discourse model it is then verbalized and subsequently uttered.

The model gives some necessary conditions for the speaker to be able to have the intention of providing the hearer with new information. The speaker can only have this intention if she can indeed assume that the information was not known to her hearer before the discourse. This is modelled in the model as a discrepancy between the speaker's primary general epistemic model which represents her own views and beliefs, and her secondary general epistemic model, which models her views about her hearer's views. There must be some information in the primary general epistemic model which is not present in the secondary general epistemic model.

Using the same terminology one can also give sufficient conditions for the speaker to have the intention to elicit new information from her hearer by asking questions. Thus, the speaker can only ask a question for new information if she has reason to believe that the hearer has the relevant information. In this case the speaker assumes that the informational content of the hearer's general epistemic model is richer than that of her own and she is asking the hearer to provide her with the information she expects him to have available. It has to be admitted that this cannot be modelled with the machinery as it stands, because a belief on the part of the speaker that the hearer knows more about something than she herself, is not something that can be modelled simply by comparing what the speaker believes and what she believes the hearer believes. Rather, it is a "meta-belief" on the speaker's part about her hearer's beliefs, and the model as it stands cannot cope with meta-beliefs.

Clearly, the explanation of the intentions of providing or eliciting new information is a very partial one. Although the model allows one to give some conditions which have to be fulfilled in order for the speaker to have these intentions, it still is unclear where her intention to speak in the first place comes from.

5.1.2. Ongoing discourses

The processes underlying the production of discourse are modelled slightly differently in the case of sentences which are not at the beginning of a new (episode of the) discourse. One can assume that in those cases the speaker need not consult her general epistemic model in order to activate the information she wants to transfer. She will have activated the information she wants to transfer in her Discourse Specific Epitemic Model already, and can therefore use the information in her discourse specific epistemic model directly. The rest of the processes will be the same.

5.2. Hearers

The processes for the hearer also differ slightly depending on whether the utterance to be processed is the first one of a new (episode of the) discourse or not. In the first case, one can assume that none of the information in the utterance has been activated before; in the latter case some of it already exists in an activated form in the hearer's mind.

5.2.1. New episodes

In the case of a new (episode of) discourse, the hearer first receives some linguistic material. He places a preliminary interpretation of this in his discourse model and consults his general epistemic model to see if he can find long-term knowledge corresponding to the material in the discourse model. If he can, he will activate it

into the discourse specific epistemic model.

5.2.2. Ongoing discourses

The case of utterances which are not the start of a new (episode of) a discourse, the processes are slightly different. We can assume that speaker and hearer will in general obey what Clark and Haviland (1977) call the given-new contract. This implies that the utterance the speaker produces contains information which the speaker will assume to be already known to the hearer and information which the speaker assumes to be new to the hearer. If this is the case, and Haviland and Clark (1974) and Clark and Haviland (1977) give good arguments that it is, then the processes underlying the understanding of a piece of discourse must be different in the case of utterances which are not the beginning of a new episode in the discourse. The hearer can be supposed to have the given information available in his discourse specific epistemic model. The need of consultating the general epistemic model is thus replaced by the need of consulting the discourse specific epistemic model in the case of given information. However, for the new information the hearer will still have to search his general epistemic model, and activate the relevant background information.

6. Conclusion

In this chapter I discussed the overall architecture of the human language processor. The model is at first sight very similar to that of Sanford and Garrod but avoids some of the problems raised in connection with it in the previous chapter. The three components in the model, the general epistemic model, the discourse specific epistemic model and the discourse model, are though of as distinct levels of knowledge activation. Each of the three partitions contains embedded models which

are used to represent the beliefs processors ascribe to their fellow interlocutors.

I briefly discussed the way knowledge is stored in the various partitions of the model. I argued that knowledge is stored around conceptual entities which are relatively large chunks of knowledge. In chapter 5. I will return to this question and expand on the brief remarks made here.

Footnotes:-

1. The discussion is limited to 'mutual belief/knowledge'. The reason is that the only type of discourse I will consider here is dyadic, i.e. involves only one speaker and only one hearer. The notion of 'common belief' (Lewis, 1969) or 'joint belief' (McCawley,1979), which is relevant to discourses involving more processors (Clark and Carlson, 1982), is a rather straightforward generalization of the notion of mutual belief knowledge. It can be accounted for along the same lines as mutual belief/knowledge is accounted for here.

2. a and b mutually know p
if a knows p, b knows p, &
a knows that b knows p,
b knows that a knows p, &
a knows that b knows that a knows p,
b knows that a knows that b knows p, &c

3. There are obviously cases where the speaker can assume that the hearer has available in an activated format information which has not been uttered in the discourse. One can use this observation to argue that the relevant distinction then is not between information which was uttered in the discourse and information from the long-term knowledge base, but rather between information which is mutually known and information which is not.

Chapter 5:- Knowledge Representation

1. Introduction

In chapter 2. I argued that a complete theory of discourse production and discourse comprehension provides both a theory about the overall architecture of the human language processor, and a theory about the specific structures it uses or constructs in discourse. In chapter 4., I provided an overall model for the human language processor. In this chapter, I will propose a theory of the specific representations used by the processor. In chapter 3. I argued that the structures used in language are identical to those used for storing knowledge. The question of the structure of the mental models used in discourse is therefore identical to the problem of knowledge representation, a question which has received ample attention in the field of AI, (for an overview, see chapter III of Barr and Feigenbaum (1981)).

I will not discuss the different knowledge representation schemes proposed in the literature in any great detail. Rather, I will give a system for knowledge representation called a Knowledge Representation Structure (KRS). I will first outline the intuitive motivations behind it. Then the grammar of KRS will be defined. Then a denotational semantics for part of this knowledge representation structure will be defined and finally, I will turn my attention to those parts of the knowledge representation language for which no model theory has been given.

2. Intuitive motivation

In section 2.1. of chapter 4. I introduced the notion of conceptual base and conceptual entity. A conceptual entity is the basic unit in the knowledge base. It consists of a

conceptual base and a set of descriptions or beliefs. The conceptual base is the focal point around which the descriptions collocate and thus ties the different beliefs together. The overall picture is of an organization of the knowledge base around conceptual entities with beliefs in them, each of which can point to other conceptual entities.

The notion in the AI literature upon which this view of knowledge representation is based is the notion of frame, which was originally introduced by Minsky (1975). Kuijpers (1975) defines the notion as follows:-

> A frame is a structure which represents knowledge about a very limited domain. A frame produces a description of the object or action in question, starting with an invariant structure common to all cases in its domain, and adding certain features according to particular observations.

The descriptors attached to a frame are called slots. One can distinguish between many different kinds of slots. Some slots just give a simple value for an attribute which an instantiation of the frame in question is assumed to have. Others contain procedures which can be called if a value for an attribute has to be determined. A third type of slot represent the fact that whenever there is an instantiation of a frame, then another object is also expected to be present. Thus, one can expect waiter-slots to be attached to the restaurant frame. I will call this third type of slot object-slots.

The slots in a frame can be seen as ordered with at one extreme properties which are true of all instances of the frame and at the other extreme default properties, properties which are assumed to be true unless there is evidence to the contrary. Frames, or prototypes, are thus not just sets of default properties as is suggested by Israel and Brachman (1984).

It is interesting to note that this way of viewing knowledge representation has become more central in AI, especially with the advent of the object-oriented style of programming underlying languages such as Smalltalk-80 (Goldberg and Robson, 1980) and LOOPS (Bobrow and Stefik, 1983) and pioneered by Hewitt (1977).

Bobrow and Winograd (1977) define a knowledge representation language KRL which is based on intuitions which are very similar to the ones outlined here. They argue against representing knowledge around sets of facts and in favour of an organization of the knowledge base around conceptual entities with beliefs attached to them. They argue that one needs many different types of conceptual entities, such as objects, relationships between objects, scenes, events, etc. I want to distinguish specifically between two different types of conceptual entity, namely mental objects and concepts. Mental objects are chunks of knowledge about particular or specific objects. Concepts are chunks of knowledge about the properties which members of certain classes of objects typically have. The distinction corresponds to the distinction between individual and generic concepts which Brachman (1979) makes in KL-ONE. It is less general however since Brachman's concepts can also be used to represent relationships between objects etc.

A knowledge base can then be seen as a set of mental objects and concepts. Since the conceptual basis of a conceptual entity is seen as a hook around which properties collocate, every conceptual entity in a knowledge base has a conceptual basis which is unique to it.

In the next section we will define a Knowledge Representation Structure (KRS) which will be used to make the ideas outlined in this section somewhat more precise. But before we turn to the formal system it is necessary to be clear about the status of the formal system. I will distinguish between KRS as a purely formal

system, KRS as a representational system with a denotational semantics, and finally KRS as a vehicle for clarifying the intuitions outlined in this section and making them more precise.

At one level, the representation language I will present can of course be seen as a purely formal system more or less on a par with the language of first-order predicate calculus. As such, it will have to meet all the requirements usually put on a formal system:- it must be clear what the basic expressions are and how they can be combined in order to construct more complex expressions. The definition of the syntax of KRS will be given in section 3.

However, it is not enough to define the syntax of KRS. For a formal language to be acceptable as a representation language we will also have to provide it with a semantics (Cf Hayes, 1977). We will have to make clear what the expressions in the language stand for. Those advocating logic for knowledge representation give as one of their main arguments for their position that their representation language does have a denotational semantics, whereas this is not the case for other knowledge representation schemes. Frame-based systems and semantic networks in general lack a semantics. An exception is Hayes (1977) who sketches a translation of Bobrow and Winograd's KRL into first-order logic. In section 4. of this chapter, I will give an alternative partial model-theory for the particular frame-based system defined here which is more direct and therefore more intuitive. I will also argue in section 5. why only a partial model-theory is possible.

Finally, the formal system I present is intended to clarify some of the intuitions outlined in the previous section and in chapter 4. about knowledge representation in the human language processor. The formal system is thus a formalization of the frame based view of the knowledge representation also

underlying Bobrow and Winograd's KRL. The terms chosen for the complex expressions defined in section 3. will reflect this intuitive background.

Given the distinction between the formal system itself, the model-theory for the formal system and the status of the formal system as clarifying intuitions about knowledge representation in the human language processor, the status of KRS vis-a-vis other frame-based knowledge representation schemes in general and KRL in particular can be made clear:- both KRS and KRL are based on the same intuitions. KRS is less expressive than a lot of other representational schemes in the sense that everything which can be expressed in KRS can also be expressed in these other schemes but not the other way around. This certainly applies to KRL. KRS however has the advantage of having at least a partial model theory, which is more direct than Hayes' treatment of KRL. As far as intuitions are concerned, KRS is based on the same intuitions as KRL and other frame-based systems. The major innovation of KRS over other frame-based knowledge representation schemes thus lies in the fact that a model-theory is defined for it.

3. A representation schema for conceptual entities

3.1. Introduction

In this section I will define a knowledge representation scheme which makes some of the intuitions outlined above more precise. The knowledge representation scheme, KRS, resembles Bobrow and Winograd's KRL, but is less expressive. I will therefore first discuss a number of restrictions on KRS. Then I will give the syntax of the formal system.

3.2. Limitations on expressive power

There will be a number of limitations in the formal definition of the notion of conceptual entity in KRS. First, although it was argued in chapter 4. that in order to represent the beliefs of other processors, one had to be able to attach embedded mental models to a conceptual base I will not take this possibility into account here. Thus, I will be able to represent only the processor's own beliefs, although it will be shown that the system can be extended to allow for embedded mental models.

A second limitation concerns the fact that although I claimed that the beliefs were ordered with at one extreme properties which every instantiation of a conceptual entity has and at the other default properties, I will only consider two possibilities here. Beliefs are either properties considered essential by the processor or default properties. Thus, the beliefs in a conceptual entity are either considered to be necessary or are characteristics which hold true only if there is no indication to the contrary.

A third limitation concerns the range of conceptual entities. I will only consider mental objects and concepts, i.e. those conceptual entities which are mental representations of objects or classes of objects. I will therefore not be able to express for example beliefs about relationships between objects. Other knowledge representations schemes such as KRL, do allow for other types of conceptual entity, and their larger expressive power is mainly the consequence of this restriction in KRS.

A final limitation concerns the fact that the only types of slots I allow are object-slots and slots ascribing simple properties. I thus do not allow for slots containing procedures for determining the value of an attribute. Again, KRL does

contain slots of this type and therefore is more expressive.

3.3. Syntax of KRS

In this section I will give formal definitions of the central notions in KRS, the notions of conceptual entity and knowledge base. In the definition of a conceptual entity I need a language for expressing beliefs. I will therefore first define a belief representation language BRL which is to play this role.

3.3.1. BRL

The definition of BRL is given in definition (1).

Definition 1
Belief Representation Language (BRL).
Vocabulary:-
The set of basic expressions of BRL is the union of
1. a set of specific individual terms, Ind
2. a set of unspecified individual terms, Var
3. a set of generic terms, Gen
4. a set of predicate letters, Pred

There is a function i from Pred to {1,2,3,..}, assigning the number of open places to a predicate. A predicate letter P such that i(P) = n will be called an n-ary predicate letter.

Terms:-
t is a term if and only if t is an element of Ind, Var, or Gen.

Grammar:-
The well-formed formulas (wffs) of BRL are defined by the following rules:-
1. if t1,..,tn are terms, and P is an n-ary predicate letter, then P(t1,..,tn) is a wff.
2. if X is an element of Gen, and t is a term, then (t ISA X) is a wff
3. if F is a wff, then not(F) is a wff.

I will use the following meta-variables:-

a,b,c for specific individual terms
x,y,z for unspecified individual terms
X,X1,X2 for generic terms
t,s,u for terms

P,Q,R for predicate letters
F,G,H for formulas

BRL contains so-called generic terms which will be used as the names for concepts. Since a concept name can be used both in ascribing a property to an individual, as in (1), which would be translated into BRL as (2), and in ascribing a property to a concept, as in (3), translated as (4), I include generic terms in BRL. In a standard logical language generic terms function as one-place predicates.

(1) Harry is a man.

(2) (HARRY ISA MAN)

(3) A man is a person.

(4) (MAN ISA PERSON)

BRL contains the predicate ISA which can appear both between specific individual terms and generic terms and between two generic terms. This fact will be reflected in the semantics where all ISA-statements will be treated in the same way. It has been argued that such a uniform treatment of ISA-links is problematic because ISA-links allow for different interpretations. A distinction is often drawn between ISA-links asserting set-membership and those asserting a subset-relationship. In (2) for example the ISA-link has to be interpreted as asserting that the referent of *Harry* is an element of the set which is the referent of *Man,* whereas in (4) a subset-relationship is asserted between the referent of *Man* and that of *Person.*

A lot of workers in semantic networks take this criticism seriously and include different types of ISA-links in their systems. So, in KL-ONE there is a distinction between the links *superc* for superconcept and *individuates.* superc-links hold between concepts, or generic concepts in KL-ONE terminology, whereas individuates-links hold between mental objects, individual concepts in KL-ONE terminology, and (generic) concepts. Mental objects inherit information from the

concepts to which they stand in the individuates relationship. Concepts in their turn may inherit information from their super-concepts. Mental objects may then inherit information which the concepts themselves have inherited from their super-concepts. We thus have a transitive inheritance relationship.

This criticism of ISA-links and the reply of including two types of ISA-links are the consequence of an extensional interpretation of one-place predicates such as MAN and ANIMAL. The problem arises because one interprets predicates as "referring" to a set of individuals. If one interprets them intensionally as "referring" to a concept, as one is urged to do by Brachman (1976,21), then the problem disappears. One can see ISA-sentences as asserting that the first argument of the ISA-link inherits the properties from the second argument. The model-theory I define gives an intensional interpretation of generic terms and thus allows a uniform treatment of ISA-links.

Apart from the above difficulty, ISA-links pose another problem. Not all information attached to a concept is inherited by the lower conceptual entity, as examples (5) and (6) illustrate.

(5) a. Horses are widespread.
b. Sir Wattie is a horse.
c. Therefore, Sir Wattie is widespread.

(6) a. Mammals are widespread.
b. The giant panda is a mammal.
c. Therefore, the giant panda is widespread.

The examples show that certain properties are not inherited by conceptual entities lower down a generalization hierarchy, conceptual entities which appear as the first argument of an ISA-link. In section 4.2.3. I will discuss the question of which properties are not inherited along ISA-links. I will also discuss a possible (semantic)

characterisation of properties of this kind. Meanwhile, the discussion will be restricted to properties which are inheritable.

The reader will have noticed that definition (1) does not give a specific language but rather a language type. What specific language one has depends on the choices one makes for Ind, Gen and Pred. In the rest of this chapter I will use the language BRL1 defined in the definition below to illustrate various points about the formal definitions.

Definition 2
An example language BRL1.
The set of basic expression of BRL1 is the union of
1. the set of specific individual terms, Ind1, {HARRY,TOM,MARY,CLARA}
2. the set of unspecified individual terms, Var1, {x1,x2,x3,...}
3. the set of generic terms, Gen1, {MAN,WOMAN,PERSON,DOG,ANIMAL}
4. the set of predicate letters, Pred1, {OWN,MARRIED,BEAT,GIVE} with
i(BROWN) = 1, i(OWN) = i(MARRIED) = i(BEAT) = 2, i(GIVE) = 3

The following are some examples of the formulas which are possible in BRL1.

(7) BROWN(CLARA)
(CLARA ISA DOG)
GAVE(CLARA,HARRY,x1)
OWN(HARRY,TOM)
(HARRY ISA MAN)
OWN(HARRY,DOG)
(MAN ISA PERSON)
not((x2 ISA ANIMAL))

In the other definitions, we need to refer to the individual terms in a formula F occurring as individual terms, Ind(F). This notion is defined as follows:-

Definition 3
Let F be a formula, then Ind(F) is defined as follows:-
1. if F is of the form P(t1,..,tn), then Ind(F) = {t1,..,tn}.
2. if F is of the form (t ISA X), then Ind(F) = {t}
3. if F is of the form not(G), then Ind(F) = Ind(G).

3.3.2. Belief representations

We will now define the notion of *belief representations.* Belief representations are

formal constructs whose intended use, as their name suggests, is the representation of the beliefs in a conceptual entity. They are the counterpart of descriptions in KRL.

A belief representation is either a straightforward ascription of a property to the conceptual entity of which the belief representation is part, or it is a description of a relationship between this conceptual entity and other conceptual entities. In the latter case, it is possible that one wants to include a number of extra conditions on whatever the conceptual entity stands in relation to. Thus, a belief representation is either a single formula, or an ordered set of formulas consisting of a formula involving the conceptual entity in which the belief representation is to occur and some other entities, together with a number of additional conditions on these other entities. In each of the other conditions the only terms which can occur are the terms occurring in the first formula.

Earlier I mentioned the possibility of slots containing procedures for calculating the value of a certain attribute. KRS does not allow for this possibility. The reason for this limitation is the expressive poverty of BRL. BRL is a purely declarative language which does not even contain conditionals. If we wanted to extend the present definition of KRS to allow for the inclusion of procedures on slots, we would have to extend BRL to make expression of procedures possible.

Definition 4
a belief representation BR is an ordered set of formulas of BRL $<F1,..,Fn>$ such that for each Fi $(0 < i < n+1)$ ind(Fi) is a subset of ind(F1) and if F1 contains a negation, then it does not contain an unspecified individual term.

Belief representations whose first formula contains an unspecified individual term are used to represent object-slots. Given our intuitive understanding of an object-slot as a representation of the expectation that whenever an instantiation of the frame is present another object is present as well, the first formula in an object-

slot cannot contain a negation.

Example (8) shows a few examples of ordered sets of formulas which are belief representations, whereas the sets of formulas in (9) are not allowed under definition (4).

(8) <OWN(HARRY,CLARA),(CLARA ISA DOG),BEAT(HARRY,CLARA)>
<MARRY(HARRY,x),(x ISA WOMAN),(x ISA PERSON)>
<(HARRY ISA ANIMAL)>

(9) <(HARRY ISA ANIMAL),(CLARA ISA ANIMAL)>
<GAVE(x,y,CLARA),(z ISA DOG)>

The sequences of formulas in (9) are not allowed as belief representations because they contain terms in the second formula in the sequences which do not occur in the first formula.

3.3.3. Conceptual entities

Given the definition of a belief representation we can now give the central definition of KRS namely the definition of the notion of a conceptual entity. Again, the construct defined is a purely formal construct whose name reflects its intended use.

Definition 5
a conceptual entity is a triple <CB,EP,DP> with
1. CB an element of Ind or Gen
2. EP a finite set of belief representations such that if F1 is the first element of a belief representation in EP, then CB is an element of ind(F1)
3. DP a finite set of belief representations such that if F1 is the first element of a belief representation in DP, then CB is an element of ind(F1)
4. EP and DP not both empty.

The first element of a conceptual entity, CB, is the conceptual base. If CB is a specific individual term, then the conceptual entity in question is a mental object; if CB is an generic term, then it is a concept. EP and DP are the sets of essential and default properties respectively. Since CB is the conceptual base around which the beliefs collocate, it has to be a term in the first element of a belief representation.

After all, the belief representation is supposed to be a representation of a belief about whatever the conceptual entity is a mental representation of.

One of the limitations in the definition of a conceptual entity is the impossibility to represent embedded mental models in conceptual entities. One can however extend the definition to allow for this possibility:- in addition to one pair of essential and default properties in a conceptual entity representing the processor's own beliefs, one can allow for the inclusion of several other such pairs, each of which can then be regarded as representing an embedded mental model.

Example (10) illustrates the definition of a conceptual entity. It defines a mental object corresponding to Harry. There are three essential beliefs, namely that Harry is a man, that Harry is a person, and that Harry is married to Clara, who is a woman. There is one default belief, namely the belief that Harry owns a dog. It will be noted that in this belief I use a variable, thus representing an object-slot.

```
(10) < HARRY,
       <  <(HARRY ISA MAN)>
          <(HARRY ISA PERSON)>
          <MARRIED(HARRY,CLARA),(CLARA ISA WOMAN)>  >
       <  <OWN(HARRY,x),(x ISA DOG)> > >
```

3.3.4. The knowledge base

We now define the notion of knowledge base. Since the conceptual base of a conceptual entity is the hook around which beliefs collocate, the conceptual base is in principle enough to identify a conceptual entity. We therefore want to make sure that in every knowledge base only one conceptual entity has a given term as its conceptual base.

Definition 6
A knowledge base (KB) is a collection of conceptual entities such that if <CB1,EP1,DP1> and <CB2,EP2,DP2> are elements of KB and CB1 = CB2, then EP1 = EP2 and DP1 = DP2

3.4. Conclusion

In this section I defined a knowledge representation scheme which makes precise the intuitions discussed in section 2 of this chapter. In particular, I defined the notions of conceptual entity and knowledge base. In the next section I will define a denotational semantics for these notions.

4. A partial denotational semantics for KRS

4.1. Introduction

In this section I will define a denotational semantics for parts of KRS. The need for a semantics for KRS arises from the fact that KRS is a representational language. Hayes (1977) characterises a representational language as one which has a semantic theory, i.e. "an account ... of how expressions in the language relate to the individuals or relationships or actions or configurations, etc., comprising the world, or worlds about which the language claims to express knowledge." Hayes claims that a formal language becomes a representational language because its expressions carry meaning and the semantic theory defines the meanings of the expressions.

The advantages of giving a denotational semantics are twofold. First, giving a denotational semantics allows one to justify the inferences which are defined for the language. A valid argument is an argument of which it is impossible that the premises are true while the conclusion is false. By defining precisely the conditions under which a sentence in the language is true, we can justify the inferences which we allow. In section 4.3. I will use the denotational semantics to argue that the

uniform treatment of ISA-links in terms of inheritance of properties is indeed justified. Thus, every argument of the form "if (t ISA X) and P(X), then p(t)" is justified because in section 4.3. I will show that whenever the premises of the argument are true, then the conclusion is true as well.

A second advantage of providing a denotational semantics for parts of KRS lies in the fact that it allows us to determine whether a given knowledge base is consistent. A set of sentences is consistent if it is possible for all sentences to be true at the same time. Again, by specifying precisely the conditions under which sentences are true, it is possible to determine whether a given knowledge base is consistent.

In section 2. I distinguished between three different ways of looking at KRS. We can regard KRS purely as a formal language, or as a representation language, or as a vehicle for clarifying intuitions about the way knowledge is stored in the human language processor. The discussion in this section is purely at the level of KRS as a representation language. None of the points made in this section should be interpreted as having any psychological relevance. The intention is to show that KRS is acceptable from a logical point of view, and there is no psychological motivation for the points made in this section whatsoever.

Hayes (1977) gives a partial semantic treatment of KRL which involves sketching a translation algorithm from KRL into multi-sorted predicate logic. Hayes's treatment has the advantage of translating KRL into a well understood and widely known logical formalism. It has the disadvantage that a lot of the intuitions underlying KRL are lost. My semantic treatment is a translation algorithm from KRS into a logical language based on the theory of arbitrary objects developed by Fine (1983,1985). It thus involves translating KRS into a less well-known formalism. The advantage however is that the logical formalism reflects the intuitions underlying

KRS more directly.

In the definition of conceptual entities I distinguished between essential and default properties. In the remainder of this section I will restrict myself to essential properties. In section 5. I will turn to default properties.

The outline of this section then is as follows:- first, I will argue that a model-theoretic treatment of KRS is indeed possible. Then I will give an intuitive introduction to the notion of arbitrary object and other notions in the theory of arbitrary objects. After that I will make the intuitions about arbitrary objects precise by defining a model-structure. I will define a language AL which contains names for arbitrary objects. I will also define an interpretation for AL. Then I will define a translation algorithm from conceptual entities into AL, thus giving an indirect denotational semantics for them.

4.2. Model-theory

In model-theory one defines a relationship between certain set-theoretic constructs called interpretations and expressions from the language one is defining the model-theory for. The intuitive meaning of this relationship is truth (or provability or warranted assertability if you are an intuitionist). As this notion pertains to propositions, this in turn presupposes that the central construct in the representational system for which the model-theory is defined is a proposition. But the central constructs in KRS are conceptual entities. It might therefore appear that one cannot define a model-theory for KRS. However, Furukawa, Takeuchi, Kunifuji, Yasukawa, Ohki and Ueda (1984) provide a way of looking at conceptual entities which would make the notion of truth applicable to KRS as well. They define a a logic-based object-oriented programming system called MANDALA. MANDALA

contains the equivalent of conceptual entities which are called Units Worlds. Furukawa et al. propose to regard Units Worlds as sets of axioms. Thus, each Units World is treated as a set of axioms. We can of course adapt this idea to KRS and regard a conceptual entity as a set of axioms with each belief representation in it corresponding to a separate axiom. This implies the problem of giving a model-theoretic semantics for KRS largely reduces to giving a model-theoretic interpretation for belief representations and the language out of which they are constructed, BRL.

4.2.1. Arbitrary objects

As said above, in the model-theory I will use the notion of 'arbitrary' or 'generic' object which was re-introduced into the philosophical literature by Kit Fine (1983,1985). Fine develops a formal system incorporating arbitrary objects, and shows how the notion can be usefully employed in the context of systems of natural deduction. Since my motivation is different from Fine's, the details of my system differ as well. I will try to point out some of the differences between the theory proposed here and Fine's original theory, but since Fine's theory is richer than mine, the reader is advised to consult Fine (1985) for a detailed exposition of Fine's system.

Fine's theory of arbitrary objects incorporates three different notions:- the notions of arbitrary object, of object-dependence, and of value assignment. Fine clarifies his notion of arbitrary objects as follows. In addition to real objects of a given kind, there is an arbitrary object of that kind. Thus associated with the set of men, there is the arbitrary man. Conversely, associated with a given arbitrary object, there is its value range, the set of real individuals which are possible "instantiations" of the arbitrary object. An arbitrary object has all generic properties which all individuals in its value range have, notion to which we will return in section 4.2.3. An arbitrary object can therefore be regarded as the prototypical object of its kind.

The notion of arbitrary object can be seen as the counterpart of concepts in KRS. Concepts can also be regarded as prototypical objects. Similarly, the real objects in Fine's theory may be regarded as counterparts of mental objects in KRS.

The second notion in Fine's theory of arbitrary objects is the notion of object-dependence. This notion is intended to capture the intuition that sometimes the real object which one takes as the instantiation of one arbitrary object depends on the real object which one chooses as the instantiation of another. Thus, if x is an arbitrary number, and $y=x.x$, then y is a dependent arbitrary object and the value one can assign to y depends on the value one assigns to x.

The relationship between an arbitrary object and another arbitrary object which is dependent on it is similar to the relationship between a concept and one of its object-slots. Just as the value associated with a dependent arbitrary object depends on the value associated with the arbitrary object it is dependent on, the object which is taken as the instantiation of an object-slot depends on the object which is the instantiation of the concept. I will use the notion of object-dependence to give an interpretation of object-slots in concepts.

In Fine's theory it is not the case that for every value assigned to an arbitrary object, there is also a value assigned to each of its dependent arbitrary objects. Let a for example be an arbitrary integer. Then the square root of a is an arbitrary object which is dependent on a. But it is obviously not the case that we can assign a value to this dependent arbitrary object for every value we assign to a.

Given the fact that I will use the notion of object-dependence for the relationship between concept and object-slot, I will use a very strong notion of object-dependence. As the discussion is limited to the essential properties of a conceptual

entity, the existence of an object-slot means that for every instantiation of the concept, there is also an instantiation of the object-slot. Condition 2 of definition 7 will reflect this.

Finally, the notion of value assignment captures the idea that arbitrary objects get real objects assigned to them as their instantiations. Since the value of one arbitrary object may be dependent on the value of another, one cannot arbitrarily assign real individuals to arbitrary objects. A value assignment tells one which individuals can be simultaneously assigned to various arbitrary objects.

The notion of value assignment bears a resemblance to the intuitive understanding of ISA-links. (t ISA X) means that t is a possible instantiation of X. In the terminology of the theory of arbitrary objects, we can say that (t ISA X) means that there is a value assignment which assign t's referent to the arbitrary object denoted by X. Since KRS also contains ISA-sentences in which both terms are generic individual terms, we need a richer notion of value assignment than Fine's original notion. Fine (1985) briefly discusses the possibility of allowing arbitrary objects to be the values of higher-level arbitrary objects but does not develop this idea any further because he does not need it for the particular application he is interested in. Given my aim of providing a model-theoretic interpretation for conceptual entities, I will use the richer notion of value assignment. Thus, I will allow lower level arbitrary objects to be instantiations of higher level arbitrary objects. In my system, the arbitrary woman can for example be an instantiation of the arbitrary human being.

ISA-links are normally used to establish a generalization hierarchy with at the top the most general concepts and at the bottom mental objects. Given that value assignments are going to be used in the semantic interpretation of ISA-sentences, The notion of value assignment I define will have to reflect the fact that ISA-links

normally establish a generalization hierarchy. Condition 7 of definition 7 ensures that this is the case.

4.2.2. A-models

Fine makes the notions of arbitrary object, value assignment, and object dependence more precise by defining an A-model. An A-model consists of a normal model for first-order predicate logic plus three extra items:- a set of arbitrary objects A, which is disjoint from the set of real individuals; a dependence relation between arbitrary objects $<$, a binary relation on the set of arbitrary objects; and a set of value assignments V, a set of partial functions from the set of arbitrary objects into the set of real individuals. All three have to meet a number of extra conditions, which follow from our intuitive understanding of the notions involved.

My notion of an A-model differs from Fine's in a number of respects. First, the set of value assignments V is not a set of partial functions from the set of arbitrary objects into the set of real individuals, but rather a set of partial functions from the set of arbitrary objects into the union of the set of real individuals and the set of arbitrary objects. This reflects the fact that we allow arbitrary objects to take lower level arbitrary objects as their values. Secondly, the extra conditions which the set of arbitrary objects A, the set of value assignments V and the dependency relation $<$ have to meet are different from the conditions they have to meet in Fine's A-models. Rather than pointing to the differences I will define my own notion of an A-model. The reader who is interested in Fine's notion is referred to Fine (1985).

Before we can formulate the conditions which A, $<$ and V have to meet, I will first define two notions which we will need, namely the notions of dependency set and value range. The dependency set of an arbitrary object a, dep(a), is defined as

$$\{b \mid b \text{ in } A \text{ and } a < b\}.$$

Thus, dep(a) is the set of all individuals a is dependent on. The value range of an arbitrary object a, VR(a), can be defined as the set

$$\{j \text{ in the union of } A \text{ and } I \mid v(a) = j \text{ for some } v \text{ in } V\}.$$

VR(a) thus consists of all arbitrary and real objects which can be instantiations of a.

We will now formulate a number of conditions on the set V of value assignments. First, we assume that if there is a value assignment v in V such that v(a) is defined, then there is a v' in V such that v' is defined for all elements in dep(a). We will call such a value assignment completed for a. The intuition is that the value assigned to a dependent arbitrary object is determined by the values assigned to the arbitrary objects it is dependent on. Thus, if a value assignment assigns a value to an arbitrary object, then it must be possible to extend it to a value assignment which assigns values to each arbitrary object it is dependent on. Condition 3 of definition 7 reflects this.

Secondly, although arbitrary objects can themselves be instantiations of higher-level arbitrary objects, we want to prevent that an arbitrary object is the instantiation of itself. Thus, regarding functions as two-place relations, every value assignment v in V has to be irreflexive (Cf condition 5 of definition 7.)

Thirdly, we want to establish a generalization hierarchy of arbitrary objects. Therefore, we want to prevent that one value assignment assigns an arbitrary object a to a (higher level) arbitrary object b whereas another value assignment assigns b to a. Thus, if v(a) = b for some v in V, then there is no v' in V such that v'(b) = a. (Cf condition 6 of definition 7.)

Fourthly, ISA-links are in general considered to be transitive. Thus, if (t ISA X) and (X ISA Y) then (t ISA Y) as well. We want our notion of value assignment to reflect this. Thus, if a is in VR(b) and another arbitrary or real object c is in VR(a), then there is v' in V such that v'(b) = c. Because of reasons which will be made clear in section 4.2.3 we also want the converse:- if a is not in VR(b) and c is in VR(a), then there is no v' in V such that v'(b) = c. (Cf condition 4 of definition 7.)

A final condition on the set of value assignments is hinted at by Fine in his brief discussion of the extended notion of value assignment used here. He writes that a natural "typing" assumption on the higher level arbitrary objects is that if one takes their values, and then the values of these values, and so on, then eventually one hits on real individuals. This typing assumption is also required because we want to use values assignments in establishing ISA-hierarchies which have mental objects at the bottom level. To formulate this requirement, let a be an arbitrary object with value range VR(a). Then define a partial ordering ord over VR(a) as follows:- if b and c in VR(a), then if there is a v in V such that v(b) = c, then ord(c,b). The typing assumption then is that the minimal elements in VR(a) under the partial ordering ord are all real individuals where a minimal element *min* is an element of VR(a) such that for no element t in VR(a) ord(t,min). (Cf condition 7 of definition 7.) ord is guaranteed to be a partial ordering (namely irreflexive and transitive) by the other conditions we have put on V.

The object-dependence relation < also has to meet a number of conditions. First, the dependence relation is irreflexive. No arbitrary object depends on itself for its instantiation. Second, the dependence relation is transitive. If a < b and b < c, then a < c, if only indirectly. Thirdly, the converse of the dependence relation is well-founded, i.e. there is no infinite sequence of arbitrary objects <a1,a2 a3,...>

such that

$$a1 < a2 < a3 < .....$$

When we start assigning individuals as instantiations to arbitrary objects, we will stop somewhere. There has to be an arbitrary object whose value can be determined independently of any other arbitrary objects. (Cf condition 1 of definition 7). It follows from these conditions that the dependence relation has to be asymmetric, i.e. that there are no two arbitrary objects a and b such that a < b and b < a [1].

A further condition on the object dependence relation is that if a < c and an arbitrary object b is in VR(c), then there is also an arbitrary object a' in VR(a) which is dependent on b The values which are assigned to a' can be determined by looking at the values a receives when c is assigned the value which is also in VR(b). Thus, if a < c and b in VR(c), then there is an arbitrary object a' < b, such that if there is a value assignment v in V with v(c) = i and v(a) = j, and there is a value assignment v' in V such that v'(b) = i, then there is a value assignment v" in V completed for a' such that v"(a') = j and v"(b) = i. In other words, we can determine the value assigned to the arbitrary object which is dependent on the less general arbitrary object by looking at the value assigned to the corresponding arbitrary object which depends on the more general arbitrary object. Condition 8 of definition 7 reflects this. An example will clarify this condition. The arbitrary man is an instantiation of the arbitrary person. Since every human has a mother, the arbitrary mother is an object which is dependent on the arbitrary person. The arbitrary man now also has a dependent arbitrary object which is an instantiation of the arbitrary mother, and could be called the arbitrary man's mother. Moreover, the values assigned to the arbitrary man's mother given a certain value assigned to the arbitrary man, is equal to the values assigned to the arbitrary mother given that that particular instantiation of man is assigned to the arbitrary person.

In definition (7), these intuitive ideas are combined in the definition of an admissible A-structure.

Definition 7
Let < I, A, <, V > be a quadruple such that I and A (the sets of real and arbitrary objects) are disjoint and not both empty, < is a binary relation on A (the dependence relation), V is a set of partial functions from A into the union of A and I (the set of value assignments)

Let a, b, c and a' in A.
For a in A, let dep(a) = {b in A | a < b}.
For a in A, let VR(a) = {j in the union of A and I| v(a)=j for some v in V}.

Then <I ,A ,< ,V> is an admissible A-structure iff
1. the relation < is irreflexive, transitive and its converse is well-founded
2. if a < b, v in V, v(b) is defined, then there is a v' in V such that v(b)=v'(b) and v'(a) is defined.
3. if v in V and v(a) is defined, then there is a v' in V such that v' is defined for all b in dep(a)
4. if a in VR(b), then VR(a) is a subset of VR(b) and if a is not in VR(b), then the intersection of VR(a) and VR(b) is empty.
5. for no v in V, v(a) is equal to a.
6. if v(a) = b for some v in V, then there is no v' in V such that v'(b) = a
7. the minimal members of the partial order (VR(a),ord) are elements of I where ord(x,y) if there is a v in V such that v(y) = x
8. if a < c and b in VR(c), then there is an a' in A such that a' < b, a' in VR(a) and if for some v in V, v(c) = i and v(a) = j and for some v' in V, v'(b) = i, then there is a v" in V such that v"(b) = i and v"(a') = j

4.2.3. Properties of arbitrary objects

Kit Fine (1983) discusses the relation between the properties of an arbitrary object and those of the real objects in its value range. He proposes the principle of generic attribution which says that an arbitrary object has all the properties which all objects in its value range have and conversely that if an arbitrary object has a certain property then all objects in its value range have it. Although this seems an intuitive principle, there are some problems with it. Consider for example the property of being a real man. Clearly, this property is true of all individual men, but not of the arbitrary man. Or, consider the property of being an arbitrary man. Clearly, this

property is true of the arbitrary man, but not of all individual men. On the basis of these considerations, Fine distinguishes between generic and classical predicates. Intuitively, a generic predicate applies to an A-object only as a representative of a class of individuals, whereas a classical predicate applies to an arbitrary object as an object in its own right. With this distinction in mind, Fine then restricts the principle of generic attribution to generic predicates. There is no corresponding principle for the classical predicates. I remarked earlier in section 3.3.1. that there are certain properties which are not inherited by conceptual entities lower down the ISA-link. We can now see that the uninherited properties are Fine's classical predicates, predicates which do not apply to the concept as a representative of a class of individuals but rather as an object in its own right.

However, even if we restrict Fine's principle of generic attribution to generic predicates, it is still too strong for our purposes. An arbitrary object is a prototypical member of a kind. It receives its properties in virtue of the fact that it is this prototypical member. Thus, it must be true that if the prototypical member has the property then all members of the kind have the property. In other words, the following weaker principle of generic attribution is true:-

$$P(a) \rightarrow (x)(x \text{ in } VR(a) \rightarrow P(x))$$

but the reverse need not be true. Suppose that by some strange coincidence which is completely unrelated to the fact that they are members of a certain kind, all members of a given kind turn out to have a certain property, then we would not always want to say that the prototypical member of the kind also has the property in question.

P may of course be a negative property, such as "is not brown". These properties will be inherited as well. After all, if the prototypical member of a kind has a negative property, and has it in virtue of the fact that it is the prototypical member

of a given kind, then all its instantiations will not have the property either. A similar argument can be used to say that if a is not in VR(b) and j is in VR(a), then j is not in VR(b). After all, if a is not an instantiation of b in virtue of the fact that it is a prototypical member of a kind, then every other member of that kind cannot be an instantiation of b either.

The existence of classical and generic predicates is caused by the fact that one can look at arbitrary objects and concepts in two slightly different ways. In the first place, one can regard concepts as representatives of a certain kind of individuals and as prototypical members of that kind. On the other hand, one can regard concepts as objects in their own right. In the object-oriented programming language LOOPS (Bobrow and Stefik, 1983), this "ambiguity" is made explicit. LOOPS includes classes, which can be seen as the counterpart of our concepts. Classes are defined as descriptions of one or more similar objects, and classes have instances, objects described by that class. But LOOPS also has metaclasses, classes whose instances are classes. When a class is regarded as an instance of a metaclass, it is regarded as an object in its own right, and the predicates which hold of a class as an instance of a metaclass can be seen as the LOOPS counterpart of classical predicates.

One can envisage extending KRS to also include classical predicates. The treatment of classical predicates I advocate is to distinguish syntactically between generic and classical predicates, rather than the LOOPS solution of allowing metaclasses. The main reason for preferring this solution to LOOPS is the fact that it is very simple to give a semantic account of classical predicates, whereas it is harder to see how to give a satisfactory semantic account of metaclasses.

For AL, the language into which we will translate conceptual entities, I will assume that all predicates are generic. In the interpretation of AL, I will therefore make use of the notion of a well-behaved set. This notion is defined in the following definition. The definition can be straightforwardly extended to a definition of well-behaved relationships.

Definition 8
Let < I, A, <, V > be an admissible A-model. Then a subset S of the union of I and A is *well-behaved* if and only if for every a in A which is also an element of S all elements in VR(a) are also in S and if a is not in S then the intersection of VR(a) and S is empty.

By restricting the interpretations of predicates in AL to well-behaved sets and relations, we make sure that the weaker principle of generic attribution holds. If a predicate is true of an arbitrary object and the arbitrary object therefore is in the extension of the predicate, then all its instantiations will be in the extension of the predicate as well and thus have the property as well.

If we restricted the interpretation of classical predicates to non-well-behaved sets as defined in definition (9), we would ensure that classical predicates are not inherited lower in generalization hierarchies. We thus see that from a semantic point of view the distinction between classical and generic predicates can be made clear.

Definition 9
Let < I, A, <, V > be an admissible A-model. Then a subset S of the union of I and A is *non-well-behaved* if and only if for every a in A which is also an element of S not all elements in VR(a) are also in S.

The language AL contains two classical predicates, namely ISA and Dep. Dep is the linguistic counterpart of the object dependence relation. It is generic in both its open places. After all, a < b, does not imply that a < c for c in VR(b), or c < b for all c in VR(a). ISA is generic in its first open place but not in its second. After all, although (a ISA b) implies that (c ISA b) for all c in VR(a), it does imply that (a

ISA c) for all c in VR(b). It is for this reason that we do not include Dep and ISA in the set of predicate letters of AL but rather introduce them syncategorematically. We can thus restrict the interpretation of predicate letters of AL, which are all generic predicates, to well-behaved sets.

4.2.4. The language AL

To interpret conceptual entities they will first be translated into a language which contains names for arbitrary objects AL. AL is defined as follows:-

Definition 10
AL contains as basic symbols:-
- a set of individual constants, C
- a set of individual variables, V
- a set of names for arbitrary objects, N
- a set of predicate letters, Pred, together with a function i which assigns to every element of Pred its number of open places.

The set of terms of AL is the union of C and N.

The well-formed formulas (wffs) of AL are
- if P is a predicate letter with i(P) = n, and t1,..,tn are terms, then P(t1,..,tn) is a wff.
- if t1 and t2 are terms, then (t1 ISA t2) is a wff.
- if a1 and a2 are elements of N, then Dep(a1,a2) is a wff.
- if F is a well-formed formula which contains an individual constant c and does not contain any occurrences of the variable x, then (x)[Fx/c] is a wff, where Fx/c is obtained from F by replacing some occurrences of c in F by x.
- if F is a well-formed formula which contains a name for an arbitrary object a and does not contain any occurrences of the variable x, then (x)arb[Fx/a] is a wff.
- if F is a wff, then (not F) is a wff.
- if F1 and F2 are wffs, then (F1 & F2) is a wff.

AL contains two different quantifiers, one for quantification over real individuals and one for quantification over arbitrary individuals. The other connectives and the existential quantifiers (Ex) and (Ex)arb are defined in terms of negation, conjunction and universal quantification in the normal way.

The semantics for AL is given in the following definition.

Definition 11
An interpretation for AL is a pair consisting of an admissible A-model <I, A, <, V> and an interpretation function f such that

if t in C, then f(t) in I.
if t in N, then f(t) in A.
if P in Pred, then f(P) is a well-behaved i(P)-place relation over the union of A and I.

for a well-formed formula F, f(F) is defined:-
if F is of the form P(t1,..,tn), then f(F) = 1 if <f(t1),..,f(tn)> in f(P) and 0 otherwise.
if F is of the form (t1 ISA t2), then f(F) = 1 if f(t1) in VR(f(t2)) and 0 otherwise.
if F is of the form Dep(t1,t2), then f(F) = 1 if f(t1) < f(t2) and 0 otherwise.
if F is of the form (x)[F'], then f(F) = 1 if for some individual constant c not in F', the formula obtained from F' by replacing all occurrences of x by c is true in all interpretations which are exactly like the current one except possibly for the interpretation assigned to c, and 0 otherwise.
if F is of the form (x)arb[F'], then f(F) = 1 if for some name for an arbitrary object a not in F', the formula obtained from F' by replacing all occurrences of x by a is true in all interpretations which are exactly like the current one except possibly for the interpretation assigned to a, and 0 otherwise.
if F is of the form (not F'), then f(F) = 1 if f(F') = 0, and 0 otherwise.
if F is of the form (F1 & F2), then f(F) = 1 if f(F1) = f(F2) = 1, and 0 otherwise.

4.2.5. The translation of KRS into AL

We are now in a position to define a translation algorithm from conceptual entities defined in KRS into sets of axioms in AL. The basic idea is to see conceptual entities as sets of axioms with every belief representation in it a separate axiom. As belief representations in conceptual entities in KRS are sets of formulas of BRL, the basic problem is to define a translation algorithm for BRL. The only exception concerns the unspecified individual terms in BRL which will treated differently depending on whether they appear in a belief representation attached to a concept or in one attached to a mental object.

In order to translate a belief representation into a formula of AL, one first has to define a basic translation function between the vocabulary of BRL and the vocabulary of AL. A basic translation function assigns an individual constant of AL to every specific individual term of BRL, a name for an arbitrary object in AL to every generic term of BRL, and an n-place predicate letter of AL to every n-place predicate letter of BRL. Given a basic translation function tr we can now define an algorithm for translating a belief representation BR attached to a conceptual entity with conceptual base X in KRS into a formula of AL:-

1. place an ampersand (&) between all formulas of BRL in BR to give BR'
2. replace every expression in BR' by the expression in AL assigned to it by tr, if any.
3. if BR' does not contain an unspecified individual term, then return BR' and stop.
4. if BR' contains an unspecified individual term y and X is a generic term, then replace y by the first variable x in Var not occurring in BR', conjoin Dep(x,a) for all names of arbitrary objects occurring in BR', and put (Ex)arb in front of the resulting formula. Set BR' to the result and go to 3.
5. if BR' contains an unspecified individual term y and X is a specific individual term, then replace y by the first variable x in Var not occurring in BR' and put (Ex) in front of the resulting formula. Set the result to BR', and go to 3.

A few examples will illustrate how the algorithm works. In 3.3.3. I used an example similar to the following to illustrate the definition of a conceptual entity in KRS. In the original example the last property was a default property. Since the present discussion is restricted to essential properties, it is now included as an essential property.

```
< HARRY,
    <  <(HARRY ISA MAN)>
       <(HARRY ISA PERSON)>
       <MARRIED(HARRY,CLARA),(CLARA ISA WOMAN)>
       <OWN(HARRY,x),(x ISA DOG)> > >
```

Let us assume that the basic translation function tr assigns h to HARRY, c to CLARA, where h and c are individual constants in AL. Moreover, tr(MAN) = man,

tr(PERSON) = person, tr(WOMAN) = woman, and tr(DOG) = dog where man, person, woman, and dog are names for arbitrary objects in AL. Finally, tr(MARRIED) = married and tr(OWN) = own where married and own are a 2-place predicates in AL. This conceptual entity can then be translated as the following set of axioms in AL.

(harry ISA man)
(harry ISA person)
married(harry,clara) & (clara ISA woman)
(Ex)(own(harry,x) & (x ISA dog))

As a further illustration, consider the following concept which represents part of someone's concept of man.

< MAN
<(MAN ISA PERSON)>
<HAS(MAN,x), (x ISA MOTHER)>>

Assuming that tr is as above, and moreover that tr(MOTHER) = mother, a name for an arbitrary object, and tr(HAS) = has, a two-place predicate, the translation of this conceptual entity in AL is the following.

(man ISA person)
(Ex)arb(has(man,x) & (x ISA mother) & Dep(x,man))

We thus see that a conceptual entity can be translated into a set of axioms in AL. Since AL has a well-defined model-theory, we have thus, in an indirect way, also provided a model-theory for conceptual entities in KRS.

4.3. ISA-links

I argued in section 3.3.1. for a uniform treatment of ISA-links in terms of property inheritance. (t ISA X) means that t has all the properties which X has. I also said that the model-theory would reflect this. In this section I will prove this assertion. Since the model-theory is restricted to the essential properties in conceptual entities,

we can prove the correctness of the claim only as far as essential properties are concerned.

In order to prove that the uniform treatment of ISA-links in terms of property inheritance is correct, we have to prove that if the translation of (t ISA X) into AL is true in an admissible A-model, and the translation of essential properties in the conceptual entity with conceptual base X is true in this A-model, then the translation of the formulas obtained by replacing X by t in every one of these belief representations is also true in this A-model. The following theorem proves this conjecture. In the definition we will make use of the notation sub(B,X,t) which denotes the result of replacing every occurrence of X in B by t.

Theorem (soundness)
let int = <Am,f> be an interpretation for AL. Let int |= P mean that P is true in int. Let BR be a belief representation attached to a conceptual entity with conceptual base X. Then
if int |= trans(t ISA X) and int |= trans(BR)
then int |= trans(subst(BR,X,t))

Proof by induction on the length of BR.

Suppose that BR contains only one formula of BRL. Assume that int |= trans(t ISA X), then int |= (tr(t) ISA tr(X)), i.e. f(tr(t)) in VR(f(tr(X))) Assume that int |= trans(BR)
1a. Let BR = <(X ISA Y)>, then trans(BR) = (tr(X) ISA tr(Y)). Thus f(tr(X)) in VR(f(tr(Y))). But then by condition 4 in Definition 7. VR(f(tr(X)) is a subset of VR(f(tr(Y))). So, f(tr(t)) in VR(f(tr(Y))) Hence int |= (tr(t) ISA tr(Y)) and hence, int |= trans(subst(BR,X,t).
1b. Let BR = <not(X ISA Y)>, then trans(BR) = not(tr(X) ISA tr(Y)). Thus f(tr(X)) not in VR(f(tr(Y))). But then by condition 4 in Definition 7. the intersection of VR(f(tr(X)) and VR(f(tr(Y))) is empty, and f(tr(t)) is therefore not in VR(f(tr(Y))). Hence, int |= not(tr(t) ISA tr(Y)) Hence, int |= trans(subst(BR,X,t)
2a. Let BR = <P(X)>, where P is a 1-place predicate letter in BRL. Then trans(BR) = tr(P)(tr(X)). Thus, f(tr(X)) in f(tr(P)). But, f(tr(P)) is a well-behaved set and therefore VR(f(tr(X)) is a subset of f(tr(P)). Hence, f(tr(t)) in f(tr(P)), int |= tr(P)(tr(t)), int |= trans(Subst(BR,X,t)) Obviously, this proof can be extended to all predicate letters in BRL.
2b. Let BR = <not(P(X))>, where P is a 1-place predicate letter in BRL. Then trans(BR) = not(tr(P)(tr(X))). Thus, f(tr(X)) not in f(tr(P)). But, f(tr(P)) is a well-behaved set and therefore the union of VR(f(tr(X)) and f(tr(P)) is empty. Hence, f(tr(t)) not in f(tr(P)), int |= not tr(P)(tr(t)) and Hence, int |=

trans(Subst(BR,X,t)) Obviously, this proof can be extended to all predicate letters in BRL.

3. Let BR = <P(X,x)> where P is a two-place predicate letter and x is an unspecified individual term. Then trans(BR) = (Ex)arb[tr(P)(a,x) & Dep(x,a)] where a = tr(X).

Case a. Let t in (t ISA X) be a generic term and tr(t) therefore stands for an arbitrary object in VR(a). Then the truth of trans(BR) and condition 8. in Definition 7. guarantee that there is an arbitrary object dependent on tr(t) which makes (Ex)arb[tr(P)(tr(t),x) & Dep(x,tr(t))] true. Hence int |= trans(Subst(BR,X,t))

Case b. Let t in (t ISA X) be a specific individual term. Then tr(t) stands for a real object in VR(a). Let b be an arbitrary object for which trans(BR) is true. Then b < a. But, there is a v in V such that v(a) = tr(t). Then by condition 2 of definition 7, there is some v' in V such that v'(a) = tr(t) and v'(b) is defined. Either v'(b) is a real object, in which case (Ex)[tr(P)(a,x)] = trans(Subst(B,X,t)) is directly true, or v'(b) is an arbitrary object, in which case the truth of trans(Subst(B,X,t)) is guaranteed by the fact that tr(f(P)) is a well-behaved set.

The induction step is straightforward.

We thus prove the correctness of the uniform treatment of ISA-links in terms of property inheritance. In section 3.3.1. I discussed the fact that such a uniform treatment of ISA-links had been criticised because ISA-links often had different meanings. Sometimes they asserted set-membership while at other times they asserted a subset-relationship. I argued that this criticism, and the normal reply of introducing different types of ISA-link, were the consequence of an extensional interpretation of one-place predicates. The 'intensional' treatment which I proposed allows one to treat ISA-links uniformly in terms of property inheritance and thus allows one to give a formal basis to the intuitions which formed the original basis for the introduction of ISA-links. It thus follows that neither the original criticism nor the normal reply to it are necessary.

4.4. Conclusion

In this section I defined a model-theory for the structures defined in section 3. thus defining a denotational semantics for an expressively poor frame-based knowledge representation scheme. I also used the model-theory to prove the correctness of the

uniform treatment of ISA-links implicit in the definition of KRS.

5. Default properties

In the previous section I gave a denotational semantics for part of KRS. I explicitly restricted the model theory to the essential properties associated with a conceptual entity. In this section, I will turn to default properties. I will distinguish between two problems with default properties, a logical and an epistemological problem. I will first discuss the logical problem and briefly sketch two possible solutions to it. Finally, I will turn to the epistemological problem.

5.1. The logical problem of defaults

What I will call the logical problem of defaults is the question whether it is possible to give a translation algorithm for a knowledge base containing concepts which have default properties associated with them into a well-understood logical formalism. In particular, given a knowledge base in which there is a conceptual entity that has a property overwriting a default property associated with a concept to which it has an ISA-link, is it possible to translate the set of formulas in this knowledge base into a well-understood logical formalism? In section 4 of this chapter, I proposed such a translation for essential properties. The question is whether something similar can be done for default properties.

In order to make the discussion more concrete, consider the following knowledge base:-

```
<ELEPHANT,
        <<(ELEPHANT ISA ANIMAL)>
         <has(ELEPHANT,x),(x ISA MOTHER),(x ISA ELEPHANT)>>
        <<grey(ELEPHANT)>
         <is-four-legged(ELEPHANT)>>>

<CLYDE,
        <<(CLYDE ISA ELEPHANT)>
         <white(CLYDE)>>
        <>>
```

Thus, in the knowledge base we have two conceptual entities, a concept ELEPHANT and a mental object CLYDE. The concept ELEPHANT has two essential properties associated with it, namely that it is an animal and that it has a mother which is also an elephant. There are also two default properties, namely that elephants are grey, and that elephants are four-legged. The mental object, CLYDE, has two essential properties associated with it, namely that Clyde is an elephant and that Clyde is white.

The translation algorithm formulated in section 4.2.5. gives the following result for the essential properties associated with the conceptual entity ELEPHANT:-

```
(elephant ISA animal)
(Ex)arb(has(elephant,x) & (x ISA mother) &
                (x ISA elephant) & Dep(x,elephant))
```

and the following result for the essential properties associated with the conceptual entity CLYDE:-

```
(clyde ISA elephant)
white(clyde)
```

For essential properties, we have the principle of generic attribution (section 4.2.3), and its semantic counterpart, the notion of a well-behaved set:- if an arbitrary object has a property and x is an instance of that arbitrary object, then x has the property as well. Formally,

P(a) -> (x)((x ISA a) -> P(x)) & (x)arb((x ISA a) -> P(x))

Using this, we can conclude from the translation of the essential properties that clyde is an animal as well, and that clyde has a mother which is an elephant.

As far as the default properties are concerned, we would like to be able to conclude from the above knowledge base that Clyde is four-legged, but not that Clyde is grey. Thus, the second default property associated with elephants should be inherited by Clyde, but the second should not. If it were, then we would be able to derive that the contradiction that Clyde was both white and grey. But the above knowledge base is not contradictory. Indeed, the whole point of making grey a default property of elephants was to allow it to be overwritten at lower concepts. The logical problem of defaults then is the problem whether it is possible to give a translation of the default properties associated with elephant in the above knowledge base which is consistent with these facts.

In the following two sub-sections, I will sketch two possible solutions to the problem. The first one proposes a relatively complex translation for default properties, the second a more complicated treatment of isa-links.

5.1.1. A first possible solution

The first possible solution to the logical problem of defaults as formulated in the previous section is based on the observation that the principle of generic attribution allows one to eliminate names for arbitrary objects from the language [2]. Without going into details, one could use the following first-order formulas as the translation for the essential properties in the example knowledge base above.

(x)(elephant(x) -> animal(x))
(x)(elephant(x) -> (Ey)(has(x,y) & mother(y) & elephant(y)))

One could try to find a similar first-order translation for default properties. If one does, then one immediately observes an important difference between essential and default properties. Note that essential properties can be treated completely "locally", i.e. in order to find a first-order translation for essential properties, one only has to take into account the property itself. But if we apply this same translation algorithm to the default property associated with ELEPHANT, then we would get

(x)(elephant(x) -> grey(x))

which would be the wrong result because we would end up with Clyde being both grey and white. Therefore a "local" treatment of defaults is impossible and that we need a "global" treatment.

Given the definition of default properties as properties which are expected to be true of all instances of the concept unless there is evidence to the contrary for a given instance, the "global" nature of defaults should come as no surprise. After all, if there is evidence to the contrary, then this evidence should be stored in the knowledge base as well. Thus, a treatment of defaults which does not take into account the state of the rest of the knowledge base is bound to fail.

Although a "local" treatment of defaults is not possible, this should not be taken to imply that in order to translate default properties into sentences of first-order logic, one has to take into account the entire knowledge base. It is possible to determine precisely which parts of the knowledge base have to be taken into account. Since defaults are inherited unless there is evidence to the contrary associated with the instances of the concept, we only have to take into account information associated with the instances in translating default properties. Thus, in the example given above, in order to translate the default property associated with ELEPHANT we only have to look at the information associated with the instances of ELEPHANT. In this

case, there is only one instance, namely CLYDE. The proposal then would be to translate defaults exactly as if they were essential properties but to mention exceptions explicitly in the antecedent. Thus, the default property associated with ELEPHANT would then be translated as:-

(x)(elephant(x) & not(x = clyde) -> grey(x))

Note that one does not have to look at the entire knowledge base when one is translating from KRS into a standard first-order language. One can make use of the fact that if there is counter evidence for a default property associated with a concept, then this will be explicitly stored with instances of this concept. Thus, by using the fact that the information which indicates possible counter-examples to default properties can be found by following ISA-links in the reverse order, we can get around the problem of the "non-local" nature of default properties and a logical treatment of defaults seems possible.

Although this approach seems to give the right results, there are two problems associated with it. First, there is the problem that finding the appropriate translation can become very cumbersome and complicated in knowledge bases where they are deep complicated hierarchies. Suppose that associated with the concept LIVING-OBJECT there is the default property "is able to reproduce itself". Then given the enormous number of instances of this concept which there might be in the knowledge base, checking whether there is counter-evidence for a particular instance becomes a very complicated and tedious process. Finding a translation for a default property such as this becomes at best a very time-consuming process.

A second problem with the above translation algorithm is the fact that it only works for static knowledge bases. Whenever a proposition is ascribing a property to a lower level conceptual entity, one should not only add the proposition in question,

but one also has to check the translation of potentially all the higher level conceptual entities the current conceptual entity is an instance of. After all, if the newly ascribed property overwrites one of the default properties of a higher-level conceptual entity, then the translation of this conceptual entity will have to be changed.

5.1.2. A second possible solution

A second possible solution to the logical problem of defaults was suggested to me by Henry Thompson (personal communication). It relies on a completely different translation of the knowledge base. The proposal is probably best introduced by way of an example. I will initially restrict myself to two-place predicates. Consider then the following knowledge base.

```
<ELEPHANT,
        <<(ELEPHANT ISA ANIMAL)>
        <<colour(ELEPHANT,grey)>
         <number-of-legs(ELEPHANT,4)>>>

<CLYDE,
        <<(CLYDE ISA ELEPHANT)>
         <colour(clyde,white)>>
        <>>
```

Under the present proposal, this knowledge base would be translated as:-

```
e-link(isa,elephant,animal)
d-link(colour,elephant,grey)
d-link(number-of-legs,elephant,4)

e-link(isa,clyde,elephant)
e-link(colour,clyde,white)
```

The translation algorithm is relatively clear-cut. The only thing to note is that e-link is used for essential properties, and d-link for default properties.

The next step is to define the three-place predicate "has-prop" which defines the notion of property inheritance. The intuitive reading of the predicate "has-prop" is something like the first argument stands in the relation denoted by its second argument to the third argument. We also define the intermediate predicate "stored" which like "has-prop" is a three-place predicate and whose intuitive meaning is that it is explicitly stored in the knowledge base that its first argument stands in the relation denoted by its second argument to the third argument.

```
(x)(y)(z)((e-link(x,y,z)  v d-link(x,y,z)) -> stored(x,y,z))

(x)(y)(z)((stored(x,y,z)  v
          (Ew)(has-prop(isa,x,w)  & e-link(w,y,z)) v
          (Ew)((u)(-(u  =  z) -> -(stored(x,y,u))  &
             has-prop(isa,x,w)  & d-link(w,y,z)) ->
          has-prop(x,y,z))
```

The second axiom reflects our intuitive understanding of property inheritance. It says:- an object x stands in the relationship y to an object z if either it is stored under x that it stands in the relationship y to z, or if it is an instance of a concept that has as one of its essential properties that it y's to z, or if it is not stored under the object that it y's to something other than z, and it is an instance of a higher-level object which has y-ing to z as one of its default properties. The qualification in the last disjunct is necessary because we would otherwise be able to derive both has-prop(colour,clyde,white) and has-prop(colour,clyde,grey) from the above knowledge base.

There are various problems associated with this proposal however. It is a matter of further research whether they seriously undermine the proposal. A first problem is that we have restricted ourselves to two-place predicates in the knowledge base and that we might get some problems if we allow predicates with any number of arguments. Also, there might be some problems with object-slots. A second problem

concerns the fact that the first-order language into which we translate presupposes a relatively complicated ontology. In order to get the semantics right, one has to allow individuals and relationships between individuals as primitive objects in the ontology. Note that the problem of giving an adequate semantics to the language which is used might get very complicated when we allow predicates with any number of open places in the knowledge base.

5.2. The epistemological problem of defaults

In the previous section I discussed the logical problem of defaults:- can we find a satisfactory proposal for translating knowledge bases containing default properties. But there is also another problem which I will call the epistemological problem. The problem is what to do if a default property has been inherited because at the time of inheritance no contradictory evidence was there, but later contradictory evidence did become available. As one wants to keep the knowledge base consistent, this means that one of the propositions has to be retracted. The solution which I sketch here is based on work on belief revision systems (also called truth or reason maintenance systems) which allow for the retraction of propositions in a knowledge base. It is beyond the scope of this thesis to provide a complete review of the literature in this area. I will only briefly discuss the basic ideas underlying this work and sketch how it might be used to give an account of default properties. The interested reader is referred to Doyle and London (1980) for an extensive bibliography.

5.2.1. Belief revision systems

Let us, for the sake of this discussion, assume that a knowledge base is simply a set of propositions. Then we can define a belief revision system as a reasoning system that can cope with retraction of some of the propositions in its knowledge base. The

problem of course is that retraction of a given proposition has repercussions for other propositions in the knowledge base. Thus, if a given proposition is retracted, then one also has to retract - or at least try to re-prove - those propositions in the knowledge base which were derived using the retracted proposition. We will call this 'forward belief revision'.

But the situation is more complicated. Suppose that because of some observation, we discover that some non-axiomatic proposition in the knowledge base has to be retracted. Then we have to look at the propositions which were used in its derivation. After all, if the conclusion of a valid argument turns out to be false, then one of the premises must be false as well. In order to maintain the consistency of the knowledge base we will have to retract the premise in question. We will call this 'backward belief revision'.

Both forward and backward belief revision have a number of consequences for the way propositions are stored in a knowledge base. In order for forward belief revision to be efficiently possible, one stores with every proposition pointers to propositions which were derived using it. If it then turns out that a given proposition changes its truth value, one can use these pointers to determine efficiently which other propositions have to be revised. For backward belief revision to be possible we store with every proposition in the knowledge base some justification, i.e. with every proposition we store some information about how it came to be included in the knowledge base. The proposition could have been an axiom, or it could have been derived using other propositions. In the latter case, we have to store pointers to the propositions which were used in its derivation. When we discover that a non-axiomatic proposition in the knowledge base has to be retracted, we can use these justifications to track down the premises used in its derivation, i.e. we can do so-

called 'dependency-directed backtracking'. (McAllister, 1980).

5.2.2. Defaults in belief revision systems

Using the ideas of belief revision systems, we can now outline a proposal for dealing with the epistemological problem of defaults. The basic idea is to store with every axiom in a belief revision system whether it is an essential property or a default property. We can then use this information to influence the belief revision process.

If we find out that a given proposition is false, then we can still use the pointers stored with it to propositions which were derived using it to determine which other propositions in the knowledge base must be revised. Thus, defaults can be treated as any other proposition as far as forward belief revision is concerned. The main change concerns backward belief revision. Let us suppose that a non-axiomatic proposition has to be retracted. Then, at least one of the proposition used in its derivation must be responsible for the fact that the proposition in question was originally derived. The point in doing backward belief revision is to blame one of these propositions, to retract the culprit and then to do backward and forward belief revision on it. In the belief revision system I am sketching here, this step will be different if the culprit turns out to be a default proposition. In this case, it is not necessary to retract the culprit. All one has to do is store the proposition which originally changed its truth value as an exception with the default proposition in question. After all, the main difference between default properties and essential properties is that the former are only true if there is no evidence to the contrary, whereas the latter are always true. The fact that one of the propositions derived from a default property turns out to have the wrong truth value, means that there is evidence to the contrary for this particular case. However, the fact that a counter-example has been found, while critical for an essential property, is not necessarily

lethal for a default property.

The following example illustrates the proposal. Suppose that (11) is a default property in some knowledge base KB, while (12) is an essential property in KB.

(11) Birds can fly.

(12) Birds have wings.

Suppose that we find out that (13).

(13) Fred is a bird

Then we can derive both (14) and (15) and with these proposition we will store some justifications, i.e. pointers to the axioms used in their derivation.

(14) Fred can fly.

(15) Fred has wings.

However, when we have a closer look at Fred, we see that Fred is a kiwi and therefore cannot fly and does not have any wings to speak of. Thus, we find out that both (14) and (15) are false. Assuming that there is no reason to doubt the truth of (13), in a standard belief revision system this would mean that one had to retract both the proposition that birds can fly and that birds have wings. In our revised belief revision system, we would have to retract only the latter proposition. After all, it was assumed to be an essential property in KB and we have just found a counter-example. Therefore, it cannot be true without qualification. The proposition that birds can fly on the other hand was supposed to be only a default property and thus is assumed to be true of birds only if there is no evidence to the contrary. The fact that we have evidence to the contrary for Fred is immaterial.

The reason we store the exceptions with default properties has to do with the fact that too many counter-examples may make one doubt the truth or usefulness of the default property. Thus, if I believe as a default that Scots are mean, then too many generous Scots will force me to revise my opinion. The question of how many counter-examples are needed is a very hard one.

5.3. Conclusion

In this section I discussed default properties in some detail. I mentioned two problems which arose out of the notion of default, a logical and an epistemological. For both problems, I outlined possible solutions. There is no doubt that more work is needed here.

6. Conclusion

In this chapter I discussed some intuitions about the way in which knowledge may be represented in the human language processor. I made these intuitions more explicit by formally defining an expressively somewhat impoverished knowledge representation scheme. I defined a model-theory for part of the knowledge representation language. I then sketched an alternative treatment for that part of the knowledge representation language for which there was no model-theory. In the next chapter I will use some of the intuitions underlying the knowledge representation scheme to discuss some expressions in natural language.

Footnotes:-

1. Suppose that $a < b$ and $b < a$. Then we can construct an infinite sequence of arbitrary objects such that

$$a1 < a2 < a3 < \ldots$$

For every odd i, let $ai = a$ and for every even i let $ai = b$.

2. Strictly speaking, the principle of generic attribution could only be used to eliminate names for arbitrary objects if it was an equivalence. However, it is one-directional for reasons discussed in 4.2.3. Therefore, we cannot eliminate names for arbitrary objects completely. However for the purposes of the discussion here, this complication is not critical.

Chapter 6:- Reference and referring expressions

1. Introduction

The system developed in chapter 4 and chapter 5 is ultimately intended to be used in a theory of discourse comprehension and discourse production. In this chapter, I use the model to sketch analyses for some expressions in English, namely indefinite NPs, definite descriptions, personal pronouns and universally quantified NPs. I will show that the idea of having various partitions in the overall model for the human language processor reflecting various degree of activation of knowledge is relevant for the distinction between definite descriptions and pronouns. I will also argue that the distinction between concepts and mental objects which was formally defined in chapter 5. is relevant for the distinction between singularly referring expressions and universally quantified expressions. But before I turn to a discussion of these various expressions, I will first discuss some of the consequences which the model has for the notion of reference.

2. Reference

The theory I put forward in chapters 4 and 5 above has a number of consequences for the analysis of reference. According to the 'traditional' or 'absolute' notion of reference which in the modern literature can be traced back to Frege (1892), the referent of a linguistic expression is an object of some sort in the world or in a model. Thus, the referents of singular referring expressions are objects in the world or model while the referents of sentences are truth values. Frege never was explicit about the referents of predicate expressions, but it was certainly within the Fregian tradition

that Carnap (1947) proposed to see the referent of a n-ary predicate, or the extension, as a n-ary relation between individuals. According to the traditional view of reference, the referent of an expression is determined once and for all, and is independent of the state of speaker and hearer. This notion of reference is captured by the interpretation function in a model for first-order logic.

Given my aim of providing a model for the processor, I am not primarily interested in the notion of reference. The central question is how a hearer grasps what the referent of an expression is on a certain occasion of use. We are interested in "how language and the world are related in the human mind, ... how the mental representation of sentences is related to the mental representation of the world." (Johnson-Laird, 1982;6-7). At least from the cognitive point of view adopted here reference always has to be seen as mediated through the mental representations discourse participants entertain of the world.

The process of referring is analysed as follows:- if a speaker wants to provide her hearer with information about an object, then this presupposes that she has activated her knowledge about this object, that she has activated one of her conceptual entities. People do not construct sentences without being aware of what they want to say. Thus, the act of referring is mediated through the mental representation the speaker entertains of the intended referent.

The process of understanding a referring expression is analysed as follows:- the hearer when interpreting an utterance in which the speaker introduces a new object into the discourse, is modelled as setting up a new discourse object. He will then try to establish a connection between it and knowledge in the long-term knowledge base. This means that he will either find an already existing conceptual entity which in some sense fits the expression the speaker used, or that he will create

a new conceptual entity. Thus, for the hearer the process of understanding a referring expression essentially involves a process of establishing a relation between the expression and a conceptual entity in his knowledge base, which either existed before the utterance was processed or came into existence as a consequence of processing the utterance. In both cases, I will use the term *cognitive access* for the relation holding between the expression and the conceptual entity.

Clearly, a speaker does not just use referring expressions to introduce new objects into the discourse. She can also use referring expressions to talk about already established discourse objects. In these cases, the hearer can access the discourse objects directly, and there is no need to establish a relationship between discourse objects and conceptual entities.

When the hearer has cognitively accessed a conceptual entity it is possible that he assumes that there is a direct relation between it and some real-world object. He can take the conceptual entity he has cognitively accessed to be the representation of some external object. If he does, he will take the speaker to be referring to this real-world object. However, whether he can determine what the real-world referent of the expression in question is, depends on whether he has enough information in the mental object he has cognitively accessed.

The following is a graphical representation of the cognitive notion of reference [1].

```
use of an expression
          |         |
          |         | cognitive access
          |         |
reference |         . conceptual entity
          |         |
          |         | correspondence
          |         |
real-world object
```

It is important to realise that whenever a referring expression is used by a speaker there is a cognitive access relation between this use of the expression and one of the speaker's conceptual entities, and also a cognitive access relation between the expression as used by the speaker and one of the hearer's conceptual entities. However, there need not necessarily be a correspondence relation between either or both of these conceptual entities and one or more real-world objects. If there is no such relation, then the expression used to access the conceptual entity also fails to refer.

My analysis of reference is not simply a reformulation of the traditional notion of reference in mental terms. The notion of conceptual entity which I use is richer than the notion of object which is used in the standard account of reference. Cognitively accessing a conceptual entity is more than establishing a relation between an expression as used on a certain occasion and an entry in the knowledge base. It involves the activation of a large chunk of knowledge. The richer notion of conceptual entity will turn out to be an advantage mainly in the analysis of definite descriptions and universally quantified NPs.

In chapter 5 I distinguished between two types of conceptual entity, namely mental objects and concepts. In section 3 of this chapter I will discuss cognitive access of mental objects, whereas I will turn to cognitive access of concepts in section 4.

3. Cognitively accessing mental objects

3.1. Introduction

In this section I will discuss three types of expression which the speaker can use to instruct her hearer to cognitively access mental objects:- non-generic or particular singular indefinite NPs, singular definite descriptions and singular personal pronouns.

3.2. Singular Indefinite Noun Phrases

The analysis of reference led to the conclusion that the reference relation has to be seen as essentially involving cognitive access of some cognitive object. Singular indefinite NPs can stand in a cognitive access relation to two different kinds of discourse objects. The discourse object can either be a mental object or a concept. In the first case, which I will call the particular use, the indefinite is interpreted as an instruction to cognitively access a specific mental object which is new in the sense that it has not been talked about before in the discourse. In the second case, the generic use, the indefinite will be interpreted as an instruction to cognitively access a concept. I will discuss particular uses in more detail in this section and I will return to generic uses in section 4.

3.2.1. Other uses of indefinites

Before I discuss particular uses in more detail, I want to make clear that the claim that in the particular use of an indefinite the speaker instructs her hearer to access cognitively a mental object does not apply to all non-generic uses of indefinites. There are a number of examples which are not covered by the analysis proposed here.

Du Bois (1980), who defends a similar thesis about indefinites, explicitly limits the discussion to the occurrences of indefinite NPs which are used to speak about an object as an object with continuous identity over time. Karttunen (1976) gives a test which can be interpreted as determining whether an indefinite is used in such a way:- an indefinite is used to speak about an object as an object only if it justifies the occurrence of a co-referential pronoun or definite description later in the text. These criteria rule out a number of non-generic uses of indefinites. One case concerns the use of an indefinite in a predicative nominal, exemplified in (1).

(1) John is a doctor.

The phrase *a doctor* is not used to speak about an object as an object, and does not allow for a subsequent co-referential definite pronoun or definite description.

Another use which is ruled out by the above criteria is illustrated in (2).

(2) Nigel and Alan went out for a pint.

The speaker does not use the indefinite *a pint* to speak about an object as an object. Moreover, subsequent co-referential pronouns or definite descriptions to refer to the pint are not allowed. Du Bois uses the term 'object incorporation' for examples of this type. His explanation is that the discourse participants treat the entire predication as a unary concept:- both speaker and hearer have a concept *going out for a pint.* There are equivalent expressions involving object incorporation which do not contain indefinites, such as *kicking the bucket* and *having sex.*

Another use which is explicitly ruled out by Du Bois and Karttunen is occurrences of indefinites within the scope of a negation, as in (3) *a car* in sentence (3) is not used to talk about an object as an object and does not allow for a subsequent occurrence of a co-referential pronoun.

(3) Bill did not buy a car.

I will however return to examples like these in 3.2.3. because the situation is more complicated than it appears at first sight.

Before we move on, it must be stressed that the test which Karttunen (1976) proposes mentions *co-referential* pronouns. It is possible to follow sentence (1)-(3) by sentences containing pronouns which have the indefinite as their antecedent. (4) would be an example for (3).

(4) They are too expensive.

Although *they* in (4) has *a car* in (3) as its antecedent, it is not co-referential. Rather, *they* refers to the type of objects of which *a car* is an example. But this relation is not one of co-referentiality and therefore does not meet the Karttunen criterion. I will ignore pronouns of this type here.

3.2.2. Understanding and producing indefinites

In a particular use of an indefinite a hearer is instructed either to create a completely new mental object in his discourse model, or to activate into his discourse model a mental object which has not been activated earlier in the discourse. In the first case, the mental object which the hearer creates in his discourse model, will be marked as an instance of the concept which has to be activated into the hearer's discourse specific epistemic model. The concept will correspond to the noun used in the indefinite. This use is most common in cases where the speaker wants to provide information to the hearer about an object the latter did not know about before. An example is:-

(5) I met a girl last night.

In the other case, the hearer is instructed to activate a mental object into the discourse. Thus, the speaker assumes that the hearer has a mental object in his General Epistemic Model and instructs him to activate it. This use is most common in cases where the speaker wants to elicit information from the hearer about this object. An example is:-

(6) You have a new girl friend, I hear. Tell me about her.

Note that the speaker could have used a definite description here. Thus, in this context (7) is almost equivalent to (6).

(7) Tell me about your new girl friend.

The main difference is that unlike in (7) in (6) the speaker first explicitly states that she knows about the existence of the hearer's new girl friend. The role of the *I hear* in (6) is presumably to check up on this information. Other examples of this use also have hedges of this type. It follows that examples of this kind are less appropriate if the existence of the object in question is mutually known to hearer and speaker. Intuitively, this prediction holds.

The main problem posed by indefinites as far as speakers are concerned are the circumstances under which the speaker uses an indefinite rather than any other "referring" expression such as a proper name or a definite. It has been claimed that the main difference between indefinites and definite descriptions is that the former are used for first mentions and the latter for second and subsequent mentions (e.g. Heim, 1982). The claim is false at least as far as definite descriptions are concerned:-. definite descriptions can be used to introduce objects into the discourse which have not been explicitly mentioned before, but only if the speaker makes some specific assumptions about the information available to the hearer in the discourse specific or the general epistemic model (Cf section 3.3.). If these conditions do not apply, then

the speaker will use an indefinite.

The analysis put forward implies that from a cognitive point of view indefinites can be used as fully referring expressions. Indefinites can be used when the speaker has cognitively accessed a mental object which she believes to stand in a correspondence relation to some object in the world. And in response to an indefinite, a hearer may create a mental object (or cognitively access an already existing mental object) which he also believes to correspond to some real world object. Fodor and Sag (1982) claim that there are a number of ways in which the speaker can signal that the indefinite is meant to be interpreted as a full referring expression. They claim that the indefinite article is lexically ambiguous between a referential and a quantifier reading. In the case of quantifier readings, the indefinite is used as an existential quantifier whereas in the referential reading the indefinite is used as a referential expression. They then give a number of factors which favour a referential reading over a quantifier reading.

Before I discuss the various ways in which a speaker can signal that she intends the indefinite to be taken as a referring expressions, I want to provide two arguments against the alleged lexical ambiguity of the indefinite. First, Fodor and Sag admit that the distinction they draw is very close to the referential attributive distinction drawn by Donnellan (1966). But the referential attributive distinction is not restricted to indefinites but can also be drawn for definite descriptions (Cf Donnellan, 1966) and proper names (Stalnaker, 1970). Hence, one is forced to conclude that there is a similar semantic ambiguity for all expressions for which the distinction can be drawn, such as for example the definite article. Second, given the importance of the resolution of the alleged ambiguity for the interpretation of utterances, one would expect the lexical ambiguity to be particular to English, and

not to exist in most other languages, no matter how closely related they are. After all, other lexical ambiguities disappear when one translates the ambiguous lexical item from one language into another. However, the indefinite article shows the same "ambiguity" in Dutch, German and French as it shows in English, thus indicating that there is good reason to believe that there is no such lexical ambiguity.

Although it is dubious that the indefinite is lexically ambiguous, Fodor and Sag's observations remain and require another explanation. We can re-analyse the alleged referential attributive analysis in terms of the amount of information the speaker has available in the conceptual entity she has cognitive accessed. If the speaker believes that she has enough information available to be able to identify the object, then we are dealing with what Fodor and Sag call a referential reading. Otherwise, we are dealing with a quantifier reading [2]. An argument for this position comes from Prince's work on a type of indefinite which Fodor and Sag argue favours a referential reading.

Prince (1981) discusses *this* indefinites. Expressions of this type are usually used in special registers, such as the one for telling jokes. Examples are:-

(8) This Irishman comes into a pub and ..

(9) There is this Irishman trying to ..

Prince (1981) argues that this type of expression is an indefinite. Firstly, occurrences of this *this* in discourse have to be replaced by the indefinite article rather than the definite. Secondly, unlike definite descriptions or demonstratives, they occur in "there is" constructions as illustrated by (9). Prince shows that there are some differences between the *a/an* indefinites and the *this* indefinites. First, when *this* indefinites are used, data collected from free speech show that there is a higher probability that the referent will be referred to again within a few clauses than when

an *a/an* indefinite was used. Thus, Prince concludes, *this* indefinites are often used to signal the introduction into the discourse of a new topic. Second, there is a difference in presuppositional behaviour. Like definite descriptions, *this* indefinites usually carry an existence presupposition in the sense that the speaker who uses one is committed to the existence of the referent of it [3]. Hence, it is not cancellable under negation. Unlike the corresponding sentence with an *a/an* indefinite, i.e. (10), sentence (11) does commit the speaker to the existence of a specific car which Bill liked but did not buy.

(10) Bill did not buy a car he liked.

(11) Bill did not buy this car he liked.

The fact that the speaker is committed to the existence of the referent, and the fact that *this* indefinites are mostly used to introduce a new topic into the discourse, support the conclusion that whenever the speaker uses a *this* indefinite, she strongly suggests to the hearer that she has more information available about the object. It is thus more likely that the indefinite will get a referential reading in the sense of Fodor and Sag.

Fodor and Sag give a number of other factors which have some effect on whether an indefinite is likely to be used "referentially". They mention a correlation with descriptive richness, i.e. the amount of information packed into the indefinite. Thus, the indefinite in (12) is more likely to be referential for the speaker than attributive.

(12) Sandy didn't see a squirrel that was
chasing its tail around the oak tree.

Indefinites of the form 'a N of mine' such as *a friend of mine* are also likely to be referential for the speaker, as in (13).

(13) A friend of mine gave me 10 pounds.

Another factor favouring a referential understanding is non-restrictive relative clauses. If an indefinite is followed by a relative clause of this type, then a referential understanding is almost certain.

(14) A student in the syntax class, who has a
Ph.D. in astrophysics, cheated on the exam.

Now, a speaker uses an expression referentially in Fodor and Sag's sense if she has attached to the mental object that she cognitively accessed when constructing the expression, knowledge which she believes enables her to identify the object. In the previous pages I have mentioned a number of ways in which the speaker can signal that she intends the indefinite to be taken referentially.

3.2.3. Indefinites and existential quantification

In formal semantics, indefinites are generally analysed as involving an existential quantifier. Thus, sentence (15) is analysed as (16) which is true just in case there is at least one dog such that Socrates owns it.

(15) Socrates owns a dog.

(16) (Ex)(dog(x) & own(Socrates,x))

Heim (1982) lists two arguments for analysing the indefinite in this way rather than as a referring expression. The notion underlying Heim's arguments is the absolute notion of reference.

For the first argument, consider (17).

(17) Socrates does not own a dog.

If *a dog* is an expression which refers to a particular object, then (17) must mean that that particular thing is such that Socrates does not own it, just as (18)

means that Socrates does not own that particular thing referred to by *Fido*.

(18) Socrates does not own Fido.

However, even though this might be one reading of (17), one can clearly use the sentence to make a stronger claim:- the claim is not that Socrates does not own some particular dog, but the claim is that Socrates does not own any dog, i.e. that for all dogs Socrates does not own them. The existential quantifier analysis generalises to this case. Thus, (19) is a logical translation of (17) which expresses this stronger claim.

(19) not(Ex)(dog(x) & own(Socrates,x)).

The second argument relies on the behaviour of indefinites in sentences which also contain universally quantified NPs. Unlike real referring expressions, the argument goes, indefinites exhibit scopal properties. Thus, (20) does not imply that every child owns the same object, something which, as (21) shows, one would expect if *a dog* was a referring expression.

(20) Every child owns a dog.

(21) Every child owns Fido.

The above arguments notwithstanding, the analysis of indefinites as involving existential quantification runs into an intuitive problem. Intuitively, indefinites carry a uniqueness implication. After an utterance of (22), which is a repetition of an earlier example, there is a strong intuition that there is exactly one dog the discourse participants are talking about.

(22) Socrates owns a dog.

It is important to make clear in what sense indefinites carry a uniqueness implication. Uses of indefinites clearly do not always imply uniqueness in the world in the sense that the predicate used to construct the expression necessarily applies to

one and only one object. Thus, sentence (22) would still be true if it turned out that Socrates owned more than one dog. Indefinites do however carry a uniqueness implication in the sense of uniqueness in the discourse. After all, it is entirely possible to continue a discourse started by (22) with a sentence containing a pronoun or an anaphoric definite description, expressions which are traditionally regarded as implying uniqueness. If we follow (22) by any of these examples, the discourse would not be called false if it should turn out that Socrates had more than one dog which bit the postman, although it might be claimed that the speaker who knows that this is the case has been misleading. Heim (1982) cites this observation as a counter-argument to the existential quantifier analysis of indefinites.

(23) It always bites the postman.

(24) The animal always bites the postman.

Seeing particular uses of indefinites as instructions to the hearer to cognitively access a mental object solves both the problem of the uniqueness intuition and the problem of personal pronouns and definite descriptions. In response to the indefinite the hearer has activated or created a unique mental object which may be accessed by means of a pronoun or a definite description, as I will argue in the next section. However, this unique mental object does not have to correspond to a unique object in the world, and uniqueness in the discourse therefore does not necessarily imply uniqueness in the world.

It remains to be seen how this analysis can deal with sentences involving negation and universally quantified sentences. In this section I will deal with the problem posed by sentences containing a negation operator, whereas in the section 4. I will deal with sentences in which there is a universally quantified NP.

In chapter 4 I introduced the notion of embedded models which could be used to represent the beliefs ascribed to other processors. Fauconnier (1979) extends what is basically the same machinery and proposes to see negation operators as expressions which the speaker uses to instruct the hearer to create an embedded negation space. This analysis implies that analysis of indefinites as instructions to the hearer to set up a new discourse object generalizes to those cases where the indefinite occurs "within the scope" of a negation operator. If the indefinite occurs in such a position, it is to be regarded as an instruction to the hearer to set up a new discourse object in an embedded negation space.

An argument which can be put forward in defence of the embedded model treatment of negation operators, is that there is a possible ambiguity in the interpretation of referring expressions in sentences containing space-creating operators. They can either be interpreted as an instruction to access cognitively an object existing in the embedded model, or as an instruction to access cognitively an object existing in the top-level model. This also is the case for sentences containing negation operators. In formal semantics terminology, an indefinite which appears "inside" a negation operator allows for both a wide scope reading, in which case it is to be seen as an instruction to cognitively access an object in the top-level model, and a narrow scope reading, in which case it is to be seen as an instruction to cognitively access an object in the embedded negation-space. The embedded model treatment makes predictions about which indefinites are more likely to be interpreted as instructions to cognitively access objects existing at the top-level, (or about which indefinites are more likely to receive wide scope). Certain indefinites were shown to be more likely to receive referential interpretations in Fodor and Sag's sense. These types of indefinites are thus more likely to receive an interpretation whereby the hearer cognitively accessed an object in the top-level model, rather than in the

embedded negation-space. Given the fact that marking an indefinite for a referential interpretation implies that the speaker has enough information available to be able to identify the object, this is what one would expect. In formal semantics terminology, the more "referentially inclined" indefinites tend to have wide scope over the negation operator. The following sentences corroborate this prediction:-

(25) I did not see a friend of mine.

(26) I did not buy this car I liked.

(27) John did not meet a man from Arkansas who
Mary had been going out with for twenty years
and who impressed everybody else by his knowledge
of pre-war British built motorbikes.

(28) Mary did not speak to a student in the syntax-class,
who has a PhD in astrophysics.

3.3. Definite descriptions

It has often been claimed that, apart from generic uses, definite descriptions have two distinct functions. They can be used to access an already activated piece of knowledge, or to instruct the hearer to activate a new piece of knowledge. The first use will be called "anaphoric uses"; the second "introductory". I will argue that there is a common factor between these two uses and that this common factor is central in the explanation of definites.

Definite descriptions can of course be used generically. The explanation of this use is very similar to that of generic indefinites which was discussed in the section 4. of this chapter. The discussion here will therefore be exclusively concerned with particular uses of definite descriptions [4].

The basic claim about particular uses of definite descriptions is very similar to the one made by Christophersen (1939). Since Hawkins (1978) explicitly bases his analysis on Christophersen as well, there is a large similarity between his analysis and mine. Christophersen wrote that a definite description used particularly, as opposed to generically, stands for a particular individual known to both speaker and hearer (Christophersen, 1939;28). Using the notion of mutual knowledge discussed in chapter 5. and the notion of knowledge activation, I want to rephrase Christophersen's claim as follows:-

> a speaker, when making a particular use of a definite description, instructs the hearer to cognitively access an activated mental object uniquely satisfying the description which the speaker assumes to be mutually known or to be derivable from activated mutual knowledge.
>
> a hearer who is faced with the task of interpreting a particular use of a definite descriptions, tries to find or derive a unique activated mental object which fits the description.

The reasoning necessary for "deriving" mental objects is of the following type. The speaker assumes that the hearer has activated some knowledge. In chapter 5, I argued that knowledge is represented in frames. Hence, whenever a processor has activated certain knowledge, there will in general be a number of slots activated as well. The objects which have to be derived in order to get an interpretation for a definite description, are instantiations of object-slots in an already activated frame. Since activated knowledge is modelled as existing in the discourse specific epistemic model, one can say that the mental object which a processor is supposed to cognitively access in response to a particular use of a definite description, must exist in the discourse specific epistemic model as mutual knowledge, whether in the form of a mental object in its own right, or as a filler for an object-slot in an activated conceptual entity.

The psychological literature is not entirely clear on the question of whether a slot-filler is present when a frame has been activated. A number of experiments have been done which can be interpreted as testing this hypothesis. If the slot-filler is present when the frame has been activated, then one would expect definite descriptions which are used as instructions to cognitively access these slot-fillers to be just as easy to comprehend as definite descriptions which are used as instructions to cognitively access explicitly introduced objects. Haviland and Clark (1974) and Clark and Haviland (1977) did a number of experiments in which they compared discourses in which an entity was explicitly introduced with discourses in which the entity was strongly implied but not explicitly introduced. Haviland and Clark compared discourses like (29) in which the existence of *the beer* was explicitly stated in the first sentence with discourse of type (30) in which the existence of *the beer* was merely implied by the first sentence.

(29) Mary unpacked the beer. The beer was warm

(30) Mary unpacked the picnic things. The beer was warm.

They found that definite descriptions referring to the entity in question took longer in the second case than in the first. The reason is that in discourses of type (30) the hearer has to establish a link between *the picnic things* and *the beer*. Haviland and Clark call this process "bridging". To the extent that bridging occurs in cases where the definite description in question is used as an instruction to cognitively access a slot in a previously activated frame, the slot-fillers are not present in the same way as explicitly introduced objects are.

However, the situation is not entirely clear-cut. Garrod and Sanford (1981) did not find a bridging effect in cases where the entity in question is considered to be a necessary part of the situation being described. In these cases, one can use an

expression referring to the entity in question without an increase in comprehension time even if the entity has not been explicitly introduced into the discourse before. Thus, if one compares (31) to (32), there is no increase in comprehension time for the case where the antecedent for the definite description *the car* was not explicitly introduced.

(31) Keith drove to London last night.
The car kept breaking down.

(32) Keith took his car to London last night.
The car kept breaking down.

The conclusion which one can draw from these findings is that if the existence of the slot-filler is strongly suggested by the frame which has been activated, then there will be no increase in comprehension time for the definite description. However, if the link between frame and slot is less strong, and the link between unpacking picnic things and beer is less strong than that between driving and cars, then there is an increase in processing time. In terms of the knowledge representation scheme of chapter 5., this can be reformulated as follows:- if the object-slot is among the essential properties, then there is no increase in processing time; if the object-slot is a default property, then bridging is called for with an increase in processing time.

Hawkins (1978,1982) rightly points out that uses of definite descriptions of the slot-filler type are highly hearer-sensitive. Thus, if the notion of a grammar has somehow been activated in the speaker's and the hearer's minds, then a use of the definite description *the deep structure* is all right when one is speaking to a transformational syntactician but would be inappropriate when one is talking to someone who cannot be assumed to know much about the subject.

Before I turn to some apparent counterexamples to the claims put forward here, I will briefly discuss the referential attributive distinction which originates with Donnellan (1966). Rather than discuss the literature produced in response to Donnellan's paper in detail, I will briefly sketch my account of the distinction and its consequences for the analysis of definite descriptions.

In section 3.2.2. I discussed Fodor and Sag's claim that the indefinite article was lexically ambiguous between a referential and a quantificational reading. I also said that they admitted that their distinction was very close to the referential attributive distinction. I rejected their claim of the lexical ambiguity of the indefinite article and proposed an alternative explanation of the observations which they use to defend their thesis, which I will now generalise to the Donnellan distinction. Roughly, a speaker uses a definite description referentially if she thinks that she has enough information available to be able to identify the real-world object corresponding to the mental object she has cognitively accessed. If this is not the case, then she uses it attributively.

Johnson-Laird and Garnham (1980) also discuss the distinction. They rightly point out that for every utterance there are two contexts:- one for the speaker and one for the hearer. It is therefore possible that the same expression is interpreted differently by speaker and hearer. Thus, although a speaker may use a definite description referentially for herself, a hearer can interpreter it referentially or attributively, and similarly for attributive uses. It follows that in general there are four possibilities for every use of a definite description. The reader is referred to Johnson-Laird and Garnham's paper for examples.

The above explanation of the Donnellan distinction is drawn in terms of the correspondence relation between the mental objects which speaker and hearer have cognitively accessed and real-world objects. However, in both the referential and the attributive use, speaker and hearer will have accessed a mental object, and the account of definite descriptions put forward in this section, which is given in terms of the cognitive access relation, therefore applies to both uses.

3.3.1. Some apparent counterexamples

There are a number of apparent counterexamples to the analysis of definite descriptions proposed above. The definite description discussed most in the philosophical literature is due to Russell (1905).

(33) The king of France is bald.

It is not difficult to find examples like it as the first sentences of newspaper articles. An example is (34) which occurred as the first sentence of an article in the *Scotsman* of 5 April 1984.

(34) The Queen is to visit the US during ..

Concentrating on sentence (34), this sentence poses two problems for the basic claim concerning definite descriptions. First, the mental object corresponding to the Queen cannot be said to be activated when the newspaper article is read. After all, if one assumes it is, then one is faced with the undesirable consequence that all information which is potentially relevant when reading a newspaper must have been activated to some degree. This would lead to an overloaded discourse specific epistemic model.

A second problem concerns the uniqueness requirement. There is more than one queen in the world and readers can be expected to be aware of this.

The first problem can be explained on the basis of the principle of retrospective updating (chapter 2.):- if the hearer's knowledge base is not as the speaker expects it to be judging by her utterance, then the hearer can retrospectively update it in order to make true the assumptions which the speaker makes. Hence, if the speaker signals that she expects the hearer to have some information available in his discourse specific epistemic model, then the hearer, if he does not have it available, can update his discourse specific epistemic model retrospectively and add the relevant information when processing the utterance. Thus, when the reader comes across the definite description in (34), he realizes that the writer acts as if she assumes that the reader had activated a particular piece of knowledge. The reader then retrospectively updates his knowledge base to make the speaker's assumption true and activates a mental object satisfying the description.

The second problem can be solved if we realize that the uniqueness is not uniqueness with respect to the entire world, but rather with respect to some pragmatically defined set. Since readers of the *Scotsman* can be assumed to be aware that normally the pragmatically defined domain is Britain, and since Britain has only one queen, uniqueness is assured. The reasons the pragmatically defined set comes to be known to the discourse participants can be of very different natures, a point to which I will return in section 3.5.

The principle of retrospective updating will also explain a use of definite descriptions which is relatively frequent in the beginning of novels and which was mentioned in connection with the discourse of Sanford and Garrod's model in chapter 3. Writers often start novels with a definite description even though they cannot

reasonably expect that the referent of the definite was known to their readers. In these cases, readers will apply the principle of retrospective updating and put a mental object corresponding to the definite description in their discourse model or discourse specific epistemic model.

The principle of retrospective updating may appear to be rather *ad hoc*. However, it is not without precedent in the literature. In chapter 2. I mentioned Seuren's principle of backwards suppletion, and Clark and Marshall (1981;24-6) use a similar principle to account for a use of definite descriptions which Hawkins (1978) called *the unavailable use*. This particular use is exemplified by examples such as (35) and (36).

(35) Bill is amazed by the fact that there is so much life on earth.

(36) The woman whom Max went out with last night was nasty to him.

Hawkins sees these examples as counter-examples to his claim that definite descriptions refer to shared sets, or shared objects, based on shared knowledge. The problem with (35) and (36) is that they can introduce information which is new to the hearer into the discourse and hence cannot be explained in terms of (previously) shared sets or objects. Hawkins' solution is to derive these sentences transformationally from deep structures containing indefinites. Hence, (35) would be derived from a deep structure which could also be realized by surface structure (37).

(37) That there is so much life on earth is a fact which Bill is amazed by.

Clark and Marshall point out that Hawkins assumes that the moment of acquisition must always chronologically precede the moment of the reference act. They claim that one can find counter-examples to Hawkins' assumption and that the condition that the hearer has a shared set available before the reference act is too strong:- the moment of the reference act can precede the acquisition of the relevant

mutual knowledge. Clark and Marshal thus use a special instance of the principle of retrospective updating of the knowledge base. But if we use the principle of retrospective updating, then there is no motivation for Hawkins' transformational treatment.

The Russellian analysis of definite descriptions is that sentences containing a definite description assert that there is exactly one object which satisfies the predicate used in the definite descriptions and that that objects also satisfies the main predicate used in the sentence. Thus, the Russellian example (33) is analysed as asserting first that there is exactly one object which is the king of France and second that that object is bald.

(33) The king of France is bald.

The Russellian analysis implies that the main difference between indefinite and definite descriptions is that the latter assert uniqueness in the sense that the speaker asserts that there is exactly one object which satisfies the descriptor. It follows that if one uses a definite description to refer to an object which is not the only one to satisfy the descriptor, then the assertion one makes is false. Strawson (1950) took exception to this point. His theory was that definite descriptions *presuppose* that the object to which the definite description is used to refer uniquely satisfies the descriptor. If there is no such object or if there is more than one, then the sentence fails to have a truth value.

The analysis I propose is essentially Strawsonian in character. The uniqueness and existence of the object which is intended as the referent of the definite description is presupposed in the sense that the speaker assumes that the hearer has activated the mental object corresponding to the intended referent. However, in cases where this assumption breaks down, the principle of retrospective updating allows the

hearer to behave in a way which is in accordance with the Russellian analysis. Thus, if there is not an object already activated which uniquely satisfies the descriptor, then the hearer can retrospectively update his knowledge base. The state of the knowledge base after the retrospective updating is the same as after the interpretation of an assertion stating that there is unique object satisfying the descriptor. It is in this sense that sometimes definite descriptions can be taken as 'asserting' uniqueness and existence.

The above remarks should not be taken to imply that the definite description can be used in two different ways. The basic analysis is Strawsonian in character. It is only when there is not unique mental object available yet that the definite description warrants a Russellian analysis. However, this is not a different use of definites but rather follows from the Strawsonian analysis and the independently motivated principle of retrospective updating.

3.4. Pronouns

The last type of expression I want to discuss in this section are third person definite singular pronouns, or pronouns for short, *he/she/it*. The claim is that speakers will in general use a pronoun to instruct their hearers to cognitively access one of the most highly activated mental objects. The main difference with definite descriptions is the degree to which the relevant knowledge is supposed to be activated. If the knowledge is activated to a very high degree, and can be modelled as existing in the discourse model, then the use of a pronoun is warranted. If the relevant knowledge is activated to a lesser degree, and has to be modelled as existing in the discourse specific epistemic model, then the use of a pronoun is less felicitous and the speaker will in general use a definite description.

In the literature on pronouns, three main principles have been used to explain the interpretation of pronouns. Pronouns were classified as deictic, as co-referential with their antecedent, or as being bound by their antecedent. The semantic interpretation of the pronoun then depends on how it is classified. In deictic uses of pronouns, the pronouns refers to some individual present in the non-linguistic context in which the pronoun is used. One can then say that the meaning of the pronoun is whatever it refers to. Clearly, this treatment of deictic uses of pronouns depends on there being a clear cut distinction between deictic and non-deictic uses of pronouns. I will argue in section 3.6. that it is hard to make this distinction precise and suggest that the explanation of pronouns put forward here allows one to account for both types of occurrences of pronouns on the basis of the same principles.

In non-deictic uses of pronouns, the interpretation of the pronoun is dependent on the interpretation of another expression in the linguistic context in which the pronoun is used. This expression is called the "antecedent". The relationship between the antecedent and the pronoun can be either one of co-reference or one of binding. In the former case, the pronoun is interpreted as referring to the same object as the antecedent refers to. In the latter case, the pronoun is regarded as a variable which is bound by its antecedent, which of course has to be a quantified NP.

Linguists working within this framework have been searching for structural criteria which determine whether in a sentence a given pronoun and a given NP can be related as pronoun and antecedent. The main problem with this approach is that the work is restricted to intra-sentential pronouns and does not say anything about inter-sentential pronouns. Given that the explanations are given in terms of the syntactic structure of the sentences in which the pronouns occur, it is not possible to

extend the explanations put forward to also account for inter-sentential pronouns.

Rather than go into the different syntactic proposals, I will concentrate on the processes taking place in the language processor when a pronoun is being used or interpreted. This will throw some light on some of the the questions which formal semanticists and linguists have struggled with. In this section I will concentrate on the interpretation of "co-referring" pronouns, while I will discuss the "bound" pronouns in section 4. I will also briefly discuss the notion of "deixis" in this section.

3.4.1. The interpretation of 'co-referring' pronouns

The convention governing the use of pronouns is that pronouns can only be used as instructions to cognitively access mutually known mental objects which have been activated to the highest degree by both hearer and speaker and are known by both interlocutors to be highly activated by the other. In terms of the model presented before, the speaker can only use pronouns as instructions to cognitively access mutually known mental objects in the discourse model.

The hearer when interpreting a pronoun, faces the problem of deciding which of the most highly activated objects is the intended one, the problem of pronoun resolution. The set of possible candidates is usually somewhat restricted because of the fact that personal pronouns carry some semantic information about number and sex/gender. Presumably, hearers go through the possible candidates in parallel. Whether a particular interpretation on the part of the hearer will be successful depends on a number of factors including the property which is attributed to the possible candidate for the interpretation of the pronoun and whether or not it is compatible with information which is already available about the discourse object. (Marslen-Wilson and others, 1981). The precise mechanisms which hearers use in

pronoun resolution is another matter which will not concern me here. [5].

The role of antecedents in the approach taken here is radically different from the role they play in other theories in which an anaphoric expression is seen as having a special relationship with its antecedent and receiving its interpretation through this relationship. Thus, in order to find the interpretation of a pronoun or an anaphoric definite description, one will first have to find its antecedent. Only then can one determine the interpretation of the pronoun.

In my account the role of the antecedent changes drastically. The antecedent can be seen as an instruction to set up a discourse object in the discourse model and to cognitively access a mental object. Since this mental object can be supposed by the discourse participants to be highly activated immediately after the introduction and for some time after, and to exist in the discourse model, the speaker can then use a pronoun to instruct her hearer to cognitively access this mental object. If some time has passed since the introduction of the discourse object in question, and it can be supposed to have disappeared into the discourse specific epistemic model, then the speaker can use an anaphoric definite description to instruct her hearer to cognitively access it. The role of the antecedent can thus be seen as an instruction to activate or re-activate a piece of knowledge which the speaker can then instruct the hearer to cognitively access by means of a pronoun. Antecedents thus play a role in the interpretation of pronouns only because they are used as instructions to activate a certain piece of knowledge which was used later in the interpretation of the pronoun. Pronoun interpretation thus is simply a matter of finding the appropriate referent in the discourse model or the discourse specific epistemic model.

3.4.2. Deixis

Another principle which is used in standard theories on pronouns is the notion of deixis:- a deictic pronoun receives its interpretation be referring to an object present in the non-linguistic context of the utterance. Standard theories rely on there being a clear distinction between anaphoric pronouns and deictic pronouns. For the way pronouns receive their interpretation differs radically depending on whether the pronoun is classified as deictic or anaphoric. However, the situation is not as clear-cut as standard theory would like.

Lyons (1979) argues that it is hard to make a clear distinction between deixis and anaphora. He argues that deixis is both ontogenetically and logically prior to anaphora. Using the notion of an intersubjective universe of discourse with a number of addresses each of which has stored under it a set of propositions in which the address occurs as a constituent, he argues that the accessibility of an address reflects the degree of salience which it currently has in the universe of discourse. Pronoun interpretation rests on the degree of salience; the more salient an object, the more likely the speaker, other things being equal, to use a pronoun. But salience is itself partly determined by recency of mention:- objects which were mentioned more recently are likely to be more salient. Recency of mention however is a deictically based notion. It has to be analysed as "relative proximity in time to the zero-point of the utterance" and thus is always determined relative to the moment at which the utterance takes place. Lyons concludes that anaphora rely on deixis and that in every anaphoric use there is a deictic element.

Lyons discusses the following example, used to console a friend who just lost his wife in a car crash.

(38) I was sorry to hear the news; I saw her only last week.

The example is an illustration of the fact that salience in the universe of discourse is determined no only by recency of mention, but also depends on other factors, such as the mental state the hearer can be expected to be in. Lyons claims that Buehler (1934) would classify the pronoun occurrence in (38) as deictic on the ground that there is no linguistic antecedent, and that Crymes (1968) would classify it as deictic on the basis that it is used to point to something in the intersubjective experience or common memory of speaker and hearer. Lyons however rejects this because

> the notion of intersubjective experience or common memory - formalisable as part of the universe-of-discourse - is the more general notion, without which anaphoric reference, as it is traditionally conceived, cannot be explained. In the last resort, there seems to be no reason to deny that the reference of *her* in the example ... is anaphoric

According to Lyons then, there is no fundamental logical difference between anaphora and deixis, and that any difference that there might be is quantitative rather than qualitative.

The entity in my theory which most closely corresponds to Lyons' intersubjective universe-of-discourse is the discourse model. If we substitute discourse model for Lyons' notion of the universe of discourse, then all his arguments still go through. The main difference is that discourse models are subjective and not intersubjective and therefore are preferable from the processor-centric position taken here. If one takes this line, then one is quite naturally led to a re-definition of deixis and anaphora in terms of what is going on in the discourse participants' minds. This line was taken by Buehler (1934) and further developed by Ehlich (1982) who defines the deictic procedure as follows. (Ehlich 1982,325; cf Bosch 1983,224).

> a linguistic instrument for achieving focusing of the hearer's attention towards a specific item which is part of the respective deictic space.

Ehlich defines the anaphoric procedure as follows:-

> a linguistic instrument for having the hearer continue (sustain) a previously established focus towards a specific item on which he has oriented his attention earlier.

Under this definition, the pronoun in example (38) has to be regarded as anaphoric and not as deictic the conclusion also drawn by Lyons. Pronouns can therefore occur without an explicit linguistic antecedent and yet not be deictic. Anaphorically used pronouns require that a piece of knowledge has been activated previously. The speaker then uses the pronoun to instruct her hearer to cognitively access this piece of knowledge. In deictic uses of pronouns, on the other hand, the speaker uses the pronoun to focus the hearer's attention on an object which had not been previously introduced into the discourse.

3.5. Different uses of definites and pronouns

In my theory speakers are modelled as using pronouns to instruct their hearers to cognitively access a mental object in the discourse model, whereas they are modelled as using definite descriptions to instruct their hearers to cognitively access a mental object in the discourse specific epistemic model. The only difference between some uses of definite descriptions and some uses of pronouns is thus the degree to which the mental object has been activated before. It follows that the various uses of definite descriptions which have been distinguished in the literature can also be found for pronouns and the other way around. In this section I will try to show that this prediction holds.

The following discussion will also illustrate another point. Processors may have various reasons for believing that something has been established as activated mutual knowledge. The distinction between the various uses of definite descriptions

and pronouns will be shown to be a consequence of the various reasons a processor may have to assume that a piece of knowledge is activated mutual knowledge.

Apart from the unavailable use mentioned 3.2.2. Hawkins (1978, 1982) lists the following other uses of definite descriptions.

1. the anaphoric use
2. the visible situation use
3. the immediate situation use
4. the larger situation use based on specific knowledge
5. the larger situation use based on general knowledge
6. the associative anaphoric use
7. the unexplained modifier use

In the anaphoric use, a definite description is used as an instruction to cognitively access an object which was explicitly introduced into the discourse beforehand. There are two slight variations of this use. In the first one, the speaker uses the actual information explicitly provided in the discourse, as in (39)

(39) I bought a car yesterday. The car was quite cheap.

In the other case, the speaker uses information which is derivable from the information which was explicitly provided, as in (40).

(40) I bought a car yesterday but today the blooming machine would not go.

In both cases, the definite descriptions are used as instructions to cognitively access a mental object which was set up earlier in the discourse in response to an explicit linguistic expression.

Another example of the anaphoric use of definite descriptions indicates that the term "anaphoric use" is maybe somewhat misleading. The following example can be found in Seuren (1985), but the type of example is fairly common in newspaper articles.

(41) Yesterday, a Swiss banker was arrested at

Heathrow airport. The 53 year old bachelor ...

The example indicates that one can use a definite description as an instruction to cognitively access a linguistically established mental object without using information which has been asserted explicitly before. The above example relies on certain background information about the typical Swiss banker. The following example which is structurally identical to (41) is far less natural, simply because one knows that the world-champion over 5000 metres is not very likely a 53 year old Swiss banker.

(42) Yesterday, an athletics champion over 5000 metres was arrested at Heathrow airport. The 53-year old Swiss banker ...

The examples (41) and (42) clarify one aspect of my account of definite descriptions. I claimed that the hearer had to find a unique activated mental object which fitted the description. The example illustrates that an object fits a description if the property used in the definite description is compatible with the information one already has about the object in question irrespective of whether it has been derived from background knowledge or from the linguistic input itself.

The anaphoric use of definite descriptions corresponds to the most natural examples involving personal pronouns.

(43) I bought a car yesterday but today it would not start.

The anaphoric use of definite descriptions and pronouns depends on the previous activation of information by linguistic means. Thus, in order for a use of a definite description or a pronoun to be classified as anaphoric, the mental object which the hearer is instructed to cognitively access has to have been activated into the discourse linguistically.

The second use of definite descriptions Hawkins distinguishes is visible situation use in which the speaker uses a definite description to refer to something present and visible in the situation shared by speaker and hearer. Consider for example the situation in which there is a screwdriver visible in the situation in which speaker and hearer find themselves. The speaker can then felicitously utter (44).

(44) Pass me the screwdriver please.

An example of the visible situation use involving a pronoun is example (45), uttered when speaker and hearer are walking along the street and someone on the other side starts drawing their attention.

(45) He is a friend of mine.

The visible situation use depends on the previous activation of knowledge because of some feature of the environment speaker and hearer find themselves in. Because speaker and hearer are aware of what is happening around them, and because they assume of each other that they are aware of the environment, certain aspects of it can be assumed to be activated and the speaker can use a definite description or a pronoun to say something to her hearer about such an object.

The difference between the anaphoric and the visible situation use on the one hand and the various other uses of definites and pronouns on the other, is that the other uses all depend on the previous activation of a piece of knowledge which is not identical to the mental object which the speaker instructs her hearer to cognitively access. In the anaphoric use and the visible situation use the mental object has itself been activated. In the various other uses the hearer has to draw some inferences in order to find the mental object. The inferences which the hearer has to draw are of a frame slot-filler type, where the frame is assumed to have been activated before. The different uses of definites and pronouns can then be shown to be a consequence of the

various ways in which the frame can be activated.

The fact that the hearer has to draw an inference in order to find the mental object in question makes it more difficult to find examples involving pronouns. The reason is that the descriptive content of definite description is much richer than that of pronouns. Since one can in principle infer an enormous wealth of information from activated background knowledge, one needs some information in order to know what to infer. However, we will see that if the number of objects which figure in the frame which is activated into the discourse specific epistemic model is small enough, pronouns can be used.

A first reason which a speaker can have for assuming that certain pieces of knowledge have been activated is the situation speaker and hearer find themselves in. The immediate situation use relies on this. The difference with the visible situation use is that the referent need not be directly visible. A common example is found on notices such as (46).

(46) Beware of the dog.

A corresponding example involving pronouns is the following. When speaker and hearer see a highly pregnant woman, then the speaker can felicitously utter (47) without any explicit mention of either the woman or the child which speaker and hearer think she is carrying.

(47) I wonder if it will be a boy or a girl.

Another reason background information can be assumed to have been activated is the general situation speaker and hearer are in. Thus, if speaker and hearer mutually know that American towns of a certain size have town-halls, then when they drive into an American town of such a size, the speaker can felicitously

utter (48), an example of the larger situation use based on general knowledge.

(48) I wonder where the town-hall is.

An example of this use involving pronouns is the following. We have general knowledge that houses have inhabitants. So, when we come past an expensive looking house, you can felicitously utter (49)

(49) I wonder how much they earn.

Another reason for assuming the activation of a piece of knowledge is the knowledge speakers and hearers have about each other. Thus, if speaker and hearer mutually know that the speaker goes to a certain bakery to get her lunch, then she can happily utter (50). Hawkins calls this use the larger situation use based on specific knowledge.

(50) I am going to the bakery.

Earlier we discussed an example of this type involving pronouns due to Lyons (1977,672). The example is repeated below.

(38) I was sorry to hear the news; I only saw her last week.

Yet another way of activating background knowledge is by explicitly mentioning the relevant knowledge in the discourse. This use underlies the anaphoric use mentioned before. However, there is a closely related but slightly different use which Hawkins calls the associative anaphoric use. In this case, it is not the intended referent which was linguistically introduced but rather the frame in which intended referent is a slot-filler. An example is (51).

(51) Somebody came past in a car. The exhaust fumes were terrible.

Examples of associative anaphoric use involving pronouns are given by Yule (1981). Thus in (52) we find that the pronoun is used as an instruction to cognitively

access the driver-slot of the (moving-) car-frame which was activated in the first part of the sentence.

(52) the car's coming up the junction and he starts to turn right.

Yule discusses more examples of occurrences in free speech of pronouns without explicit linguistic antecedents. One of his other examples is (53).

(53) well I saw a demolition order there actually - a few months ago - they said they were going to demolish some of the flats - which is a pity - I don't know what they're doing with Edinburgh though - as long as they don't do what they did with Glasgow.

On the basis of this and other examples of antecedentless pronouns, Yule (1982,319) claims that it is entirely possible that hearers sometimes do not actually expend any effort working out the referents of pronouns. The underlying principle might be that if a speaker signals in her utterance that the referent of a pronoun is less important that the predicated information, then the hearer will not spend much effort working out the referent. What hearers do on these occasions is interpreting the speaker's message in terms of some information which the speaker marks for attention, information which is predicated of some individual or group whose referential identity is not at stake. Thus in (52) the reference of *they* is not relevant. The message the speaker wants to convey can be understood without resolving the referent of it.

There are two general points to be made about Yule's remarks. In the first place, and Yule would certainly agree with this, there are certainly cases of antecedentless pronouns where the referential identity of the referent of the pronoun is an issue which has to be resolved. Some of the above examples illustrate this point. For these cases, one needs something along the lines of my account. Secondly, even if we admit that on certain occasions hearers do not try to determine what a pronoun is used to refer to, the question remains what prompted the speaker to use a pronoun in

the first place. Even if the hearer need not resolve the referent of a pronoun in order to be able to understand the speaker's utterance, one still needs the notion of knowledge activation because it still is the case that the speaker can only use a pronoun to refer to the object which is the most highly activated in her own mind. (Cf also Sanford, Garrod, Lucas and Henderson, 1983).

Another use which Hawkins distinguishes, the so-called unexplanatory modifier uses, is illustrated in (54). Clark and Marshall (1981,23) correctly observe that these use are what Donnellan (1966) has called attributive uses.

(54) The winner of the 1988 Olympic marathon ..

Donnellan illustrates the attributive use in a situation where Smith is lying on the floor brutally murdered. The speaker who uses the definite description *Smith's murderer* to refer to whoever murdered Smith is said to have used it attributively. However, in the very same situation one could as well have used a pronoun and express exactly the same information. Thus in the situation described above, one might as well have used (55) without any risk of misunderstanding.

(55) He must be insane.

In this section I argued for my analyses of 'co-referential' pronouns and definite descriptions. I argued that if my analyses are correct, and the main difference between these two types of expressions concerns the degree to which the knowledge which is cognitively accessed has been activated before, then for every example of a type of use of definite descriptions, there must be a corresponding example for pronouns, and the other way around. The above observations support this claim. Also, I showed that the different uses of pronouns and definite descriptions can be accounted for as the consequence of the different reasons one can have for assuming that a certain piece of knowledge has been activated.

3.6. Conclusion

In this section I discussed the expressions the hearer has available to instruct her hearer to cognitively access mental objects. The speaker can use an indefinite to introduce into the discourse a new mental objects, i.e. a mental object which was not activated before in the discourse. I argued that a speaker can use either a definite description or a pronoun to instruct her hearer to access an already established mental object. Which expression is most felicitously used depends on the degree to which the intended mental object can be assumed to be activated. I showed that one of the predictions of this treatment, namely that for every use of a definite description there is a corresponding pronoun use, and the other way around, was indeed borne out by the facts. I also discussed some consequences of the analyses for linguistic theory.

4. Cognitively accessing concepts

4.1. Introduction

Although a lot of work has been done in procedural semantics on the type of expressions discussed in the previous section, much less has been done on the expressions which I will discuss in this section, namely generics and universally quantified NPs. I will argue that these expressions are used by speakers to instruct their hearers to access concepts cognitively. The basic claim of this section then is that generic indefinites and definites and universally quantified NPs of the sort exemplified in (56) are used to instruct the hearer to access cognitively a concept.

(56) all boys ..
many boys ..
few boys ..
no boys ..

The following discussion is restricted to generics and quantified NPs of the sort exemplified in (56). I will not say anything about other expressions which are usually regarded as quantifiers such as the numerical quantifiers *one, two, three* etc. and the plural indefinite article *some*, or the quantifiers exemplified in (57).

(57) all (of) the boys ..
many of the boys ..
few of the boys ..
none of the boys ..

The reason for these restrictions is that a discussion of these various quantified expressions would have to rely on a previous discussion of plural discourse entities, something which is outside the scope of this thesis. The quantifiers are either used to introduce plural objects into the discourse, e.g. *some*, or they are anaphoric to plural discourse entities and thus rely on a previous introduction of such an entity.

A final restriction is that the only occurrences of quantified or generic noun phrases which will be considered are those in subject position.

In procedural semantics, language is primarily regarded as a vehicle for the transfer of information from speaker to hearer. In order for it to be a successful medium of transfer, language must in principle enable its users to transfer large parts of their knowledge bases. For most types of knowledge language therefore can be expected to provide the means by which processors can transfer knowledge of that type. A large part of our knowledge is of a general type. In the previous chapter, I used the term generic knowledge. Given the fact that language is likely to provide its users with the means to transfer almost any type of knowledge, and given the existence of generic knowledge, it is very natural to expect that language will enable its users to transfer generic knowledge to each other. The basic claim of this section is

that the means language provides for transferring information regarding concepts, knowledge about classes of individuals, are generic and quantified noun phrases. Generics and quantified NPs are instructions to the hearer to cognitively access a concept and attach the predication made in the sentence to the activated or newly created concept.

The explanation of the difference between the different quantified NPs crucially depends on the existence of a hierarchy in the properties attached to a concept. As argued in chapter 5, the properties attached to a concept are ordered along an axis with at the top level essential properties and lower down default properties. The claim is that every universally quantified or generic NP is an instruction to cognitively access the concept corresponding to the noun following the quantifier and that the actual quantifier determines where in the hierarchy the property mentioned in the predication of the sentence is to be attached. In the case of the universal quantifiers like *every* and the equivalent quantifier *all* the property predicated is to be attached among the necessary properties. *many, most* and equivalent quantifiers are used for default properties. *few* is used for properties which by default the frame does not have attached to it. (Cf Reiter 1980,82-83 for a similar proposal for the non-standard quantifiers). Finally, *no* is used to predicate properties which the frame necessarily does not have attached to it. Thus, the information which is transferred by (58) is that a property which according to the speaker is necessarily attached to the car-frame is that there are sparkplugs.

(58) Every car has sparkplugs.

In (59) the transferred information is that the default value for cars is having sparkplugs, in (60) that the default value for cars is not having sparkplugs, and finally in (61) that a necessary property for cars is having no sparkplugs.

(59) Many cars have sparkplugs.

(60) Few cars have sparkplugs

(61) No car has sparkplugs.

The analysis of universally quantified NPs generalises to generic uses of definite and indefinite NPs. This is an intuitively appealing result. Thus, generics such as (62) and (63) instruct hearers to attach the property of having sparkplugs to the car-frame.

(62) A car has sparkplugs.

(63) Cars have sparkplugs.

Unlike utterances involving "overt" quantifiers, when a speaker uses a generic NP, she does not give the hearer explicit instructions as to whether the property which is predicated of the concept is to be attached as a necessary property or as a default value. This is in accordance with the fact that generic sentences often are considered to be ambiguous or vague. Thus, (64) and (65) are often claimed to be "less precise" versions of (66) and (67). [6].

(64) A dog is a mammal.

(65) A dog has four legs.

(66) All dogs are mammals.

(67) Most dogs have four legs.

The explanation for this phenomenon in the present theory would be that generic sentences are essentially non-committal about where the predication made in the sentence is to be attached in the hierarchy hanging off the concept denoted by the generically used NP [7].

The predication made in quantified sentences need not necessarily involve "atomic" properties, such as *is white*. One can also use transitive verbs or bi-transitive verbs, in which cases the predication involves one or two other NPs. An example is (68)

(68) Everyone has a mother.

The treatment of this sentence in the present theory naturally follows from the treatment of quantified NPs and that of indefinite NPs proposed in section 3.2. Thus, in response to *everyone* in sentence (68), the hearer will cognitively access his person-concept. He will realize that the property expressed in the predication *has a mother* will have to be attached among the necessary properties. However, the predication is not "atomic" but contains an indefinite NP and therefore requires further analysis. I argued earlier that the speaker uses indefinites to instruct her hearer to cognitively access new mental objects. In most cases this boils down to an instruction to the hearer to create a new mental object. If everything goes according to plan in (68), then this is what the hearer will do when he is processing *has a mother*. However, unlike most of the examples discussed in the section on particular uses of indefinites, the mental object is not created independently of everything else. Rather, it is created as part of a property which is to be attached among the necessary property hanging off the hearer's person-concept. Thus, in response to sentence (68), the hearer will cognitively access his person-concept and create a mental object which is dependent on the person-concept. In other words, he will create a new object-slot to be attached to his person frame.

Some support for this analysis can be derived from the following observations. If there is a strong expectation that something is an instantiation of a frame which has a slot of a particular sort attached to it, then sentences like (69) in

which the speaker uses a NP with a possessive pronoun to get at the slot in question are quite natural, also in discourse initial position.

(69) John loves his mother.

However, if this is the correct explanation cases where such a connection is unexpected are harder to interpret. Clearly, they are not impossible to interpret because of the principle of retrospective updating:- one can always attach the slot to the relevant frame in response to such a sentence, or adapt one's knowledge base in another way. Thus in response to (70), which definitely is stranger than (69), the hearer can either attach an elephant-slot to the frame the vicar is an instantiation of, or set up a specific mental object as the mental representation of the particular elephant the vicar owns.

(70) The vicar lost his elephant

The prediction is that if in preceding discourse one explicitly instructs the reader to attach an elephant to the frame the vicar is an instantiation of, then the strangeness disappears. This prediction is intuitively satisfied. Discourse (71) illustrates this.

(71) Everybody in the village recently got an elephant from the industrialist who has been on Safari to Africa. The vicar got his elephant on Monday, whereas the verger got his elephant on Tuesday.

4.2. The interpretation of bound pronouns

In section 3.4. I discussed those pronoun occurrences which in standard theories are explained as involving co-reference between the pronoun and its antecedent. There is, however, a class of examples in the literature which are standardly explained on the model of variable-binding in first-order predicate calculus. I will turn to these now

and present an alternative explanation, which has the advantage that it treats these pronoun occurrences as not essentially different from the co-referring pronoun occurrences.

According to the binding explanation some pronoun occurrences have to be treated similar to variables in predicate calculus. If the antecedent of a pronoun is a quantified expression then it is regarded as binding that pronoun occurrence. When the sentence is translated into some logical language, the pronoun turns into a variable which occurs within the scope of the translation of the antecedent and is bound by it.

Examples of bound occurrences of pronouns then are (72) and (73) [8]. Example (74), which is similar in a number of respects, is usually treated as a pronoun of laziness.

(72) Every dog loves somebody who loves it.

(73) Everybody believes that he is ill.

(74) Everybody has a mother and everyone trusts her.

Adopting the system developed in Montague (1973) (72) and (73) are translated as (75) and (76) respectively.

(75) (x)(Ey)(love(x,y) & love(y,x))

(76) (x)(believe(x,^ill(x)))

One sees that in both cases the pronoun has been translated as a variable which is bound by the translation of the antecedent, *everybody*. It is not clear how one is to account for the pronoun in sentence (74). One cannot regard it as being bound by the universal quantifier, nor by the indefinite in the first sentence *a mother*, because this would imply that mother would receive wide scope over the universal

quantified NP, implying that everyone had at least one mother in common.

My explanation of occurrences of bound pronouns relies on the notion of knowledge activation and the treatment of generic and quantified expressions proposed earlier. In response to a generic or quantified expression the hearer activates a concept and interprets the predication as an instruction to attach a new property to this concept. Bound pronouns are used to instruct the hearer that the property to be attached somehow involves possible instantiations of the activated concept. It is for example possible to use pronouns to put some special conditions on object-slots which say that the slot filler has to stand in a certain relation to the instantiation of the frame. (72) is an example of this use. The information transferred is that one has to attach to one's dog-frame an object-slot such that an instantiation of the frame stands in the love relation to a filler for that slot and moreover such that the filler for that slot stands in the love-relation to the instantiation of the frame. Using KRS, the knowledge representation scheme developed in chapter 5, in response to (72) the conceptual entity *dog* will have the following belief representation added amongst its essential property. I assume that DOG is the conceptual base of this conceptual entity:-

<LOVE(DOG,x),(x ISA PERSON),LOVE(x,DOG)>

Other examples involve the idea of embedded models introduced in section 2.3. (73) is an example. The predication contains a 'space-creating' operator and the property which is to be attached to the concept involves an embedded model. Personal pronouns can then be used to instruct a hearer to cognitively access possible instantiations of the frame, which can be assumed to exist in these embedded mental models. A hearer faced with (73) is thus instructed to attach an embedded model to his person frame which contains the information that the instantiation of the frame is

ill. Using an ad hoc extension to KRS for embedded models, (73) gives rise to the following belief representation. which would be added to the rich-person-concept. Again, RICH-PERSON is assumed to be the conceptual base of the conceptual entity.

<BELIEVE(RICH-PERSON,
<<ILL(RICH-PERSON)>>)>

Example (74), finally, can be accounted for given the treatment of multiply quantified sentences of section 4.2.:- in response to the first clause of the conjunction the hearer will have activated his person-concept and attached a mother slot to it. As a consequence, both the person frame and the newly created mother slot will be highly activated. The pronoun then is used as an instruction to the hearer to cognitively access the mother-slot. Thus, in response to the first conjunct in (74), the following belief representation is added to the essential properties of the person-concept.

<HAS(PERSON,x),(x ISA MOTHER)>

This mother slot is then available as the referent of *her* in the second conjunct and in response to it the above belief representation is changed into

<HAS(PERSON,x),(x ISA MOTHER),TRUST(PERSON,x)>

The analysis of bound pronouns predicts corresponding cases involving generic uses of indefinites and definite descriptions. After all, the analysis I proposed for generic NPs is very similar to that for quantified NPs. The prediction is born out by the facts.

(77) A dog always loves somebody who loves it.

(78) A rich man always thinks he is ill.

it in (77) and *he* in (78) have to be analysed as receiving their interpretation through a mechanism similar to those (72) and in (73). Since it is difficult to see how

generics can be treated as quantifiers but have to be treated as terms referring to kinds (Carlson,1977), the solution I propose for dealing with the so-called "bound" variables is more general than the bound variable solution.

4.3. Donkeys revisited

In section 4.1. of chapter 3 I briefly discussed the so-called donkey sentences, of which (79) is an example. Donkey sentences play an important role in those theories of the mental representation of discourse which are based on truth conditional semantics. In this section I will briefly sketch an alternative treatment of donkey sentences.

(79) If a donkey is hard-working, it'll be fed.

The problem posed by donkey sentences is the fact that indefinites occurring in the antecedent of a conditional, henceforward 'donkey' positions, get a universal import. Whether the indefinite should get full universal import, as Kamp (1981) argues, or merely generic import, as Seuren (1985, 373-5) claims, is a bone of contention. My intuitions are in agreement with Seuren's and before I sketch a treatment of donkey sentences, I will put forward a few observations which show that certainly not every indefinite in a 'donkey-position' can receive a 'donkey interpretation'. What these observations are intended to show is that indefinites do not get their 'universal' import purely by appearing in a 'donkey' position but that other factors play a role as well. As some of these other factors are also important in determining whether an indefinite can get a generic interpretation, these observations are also intended as arguments for the intuition that indefinites in 'donkey' positions should be given a generic rather than a full universal interpretation.

First, indefinites which are strongly marked as favouring a referential interpretation in the Donnellan sense (cf section 3.2.2.) can only with great difficulty appear in 'donkey' positions, and if they do, they do not get a 'donkey' interpretation. The indefinites in (80) and (81) show this. But indefinites which are marked as favouring a referential interpretation cannot be given a generic interpretation either.

(80) If a donkey in the field, who has been working in the Foreign Legion, works hard, it'll be fed.

(81) If a donkey of mine works hard, I feed it.

Another observations relies on the fact that whether an indefinite in subject position can receive a generic interpretation is correlated with the tense and aspect of the verb. Thus, in (82) the indefinite can be given a generic interpretation, whereas in (83) this is much more difficult.

(82) A donkey works hard.

(83) A donkey had worked hard.

But if we form conditionals (84) and (85) with (82) and (83) as antecedents, then we see that (84) allows a 'donkey' reading whereas it is much harder to get a 'donkey' reading for (85).

(84) If a donkey works hard, it will not get its owner into trouble.

(85) If a donkey had worked hard, it would not have gotten its owner into trouble.

If we accept that indefinites which receive 'donkey' interpretations are indeed generic, then donkey sentences are no longer a special case. The treatment of generics which I proposed in the previous section applies to donkey sentences as well. A 'donkey' sentence is an instruction to the hearer to attach a new property to one of his concepts. Thus, in response to sentence (79) a hearer will add the belief representation to his donkey-concept that if an instantiation is hard-working, then it

will be fed. If we extend the notation developed in chapter 5 to include conditionals, then in response to (79), the hearer should add the following belief representation to his donkey-concept.

<(IF (HARDWORKING DONKEY) THEN (IS-FED DONKEY))>

Because the indefinite is generic, the speaker is not entirely clear whether this belief representation should be added to the essential properties or to the default properties associated with the donkey-concept.

The reader will recall that I treat the generic term DONKEY as a denoting term in the semantics of KRS. It denotes an arbitrary object. If KRS was extended to allow for conditional belief representations as well, and the denotational semantics for KRS was extended to cope with this, then the denotational interpretation of the above belief representation would be reasonably straightforward:- it would be true if the arbitrary donkey had the property of being fed if it was hard-working. Assuming that this is a necessary property and using the principle of generic attribution introduced in section 4.2.3. of chapter 4, this would entail that all individual donkeys had this property as well. Under the assumption then that the property attributed in a 'donkey' sentence is a necessary property, we arrive at the truth-conditional reading of donkey sentences which was favoured by Kamp (1981).

4.4. Conclusion

In this section I proposed to see universally quantified NPs as instructions to the hearer to cognitively access a concept. I made use of the proposals for knowledge representation made in chapter 5. To the extent that the present proposal for the mental representation of universally quantified NPs is successful, it can also be seen as an argument for the particular form of knowledge representation I put forward.

5. Conclusion

In this chapter I have re-analysed the notion of reference and argued that reference always has to be seen as involving the cognitive access of certain entities in the speaker's and the hearer's mind. I also provided analyses of the use and function of various referring expressions and of quantifiers. In these analyses I made use both of the model for the human language processing mechanism and especially the notion of knowledge activation proposed in chapter 4., and the more detailed proposals about knowledge representation in chapter 5.

Footnotes:-

1. I will completely ignore the question what objects referring expressions in novels etc. refer to. It might be possible to analyse these expressions completely in terms of the conceptual entities existing in the speaker's and hearer's minds. If this is the case, then these expressions do not refer at all. If such an analysis turned out to be impossible however, then the only alternative which remains open would be to define the notion of 'real-world' object so that it includes objects of this kind.

2. The situation is more complicated. As Johnson-Laird and Garnham (1980) point out, there are two different contexts for every utterance, one for the speaker and one for the hearer. It is therefore possible that the speaker intends different interpretations for a given expression for herself than for her hearer. I will ignore this complication here.

3. Prince observes that the existence presuppositions can be cancelled under at least certain verbs of propositional attitude. Thus, 1'. is fine and certainly does not commit the speaker to the existence of the big Glaswegian in question.
1'. Nigel says that this big Glaswegian
smashed his face, but I think he was
drunk and fell off his bike.

4. There are subtle differences between generic uses of definites and generic uses of indefinites, but I will not discuss them here (see e.g. Hawkins, 1978)

5. There is a similarity between this explanation of pronoun resolution and the focussing approach of Sidner (1983) and Grosz (1981). After all, one can regard material which is present in the discourse model as being in focus. However, unlike Sidner and Grosz, my proposal does not rely on the sort of structural and formal rules that they propose. The basic assumption is rather that the information which is already known about the possible candidates, because of previous discourse or because of activated background knowledge, is critical.

6. As Phil Johnson-Laird pointed out to me universally quantified sentences often show a similar vagueness. An example is
Everybody is wearing pink this season

7. I will ignore the differences between generically used indefinites, generically used definite descriptions, and generically used bare plurals.

8. In what follows I will not consider other examples involving quantified expressions as antecedents such as the ones exemplified below. The main reason is that they involve plural pronouns. I would like to point out though that the fact that these sentence (1') allows for a reading which makes it synonymous to (77) casts some doubt on the generality of the binding explanation. This doubt is reinforced by the fact that in what have to be pure binding cases singular pronouns are impossible with quantified expressions whose grammatical number is plural, as (2') exemplifies.
(1') Every dog loves somebody who loves them.
(2') Most dogs love somebody who loves them.
(3') * Most dogs love somebody who loves him.

Chapter 7:- Conclusion

In the previous six chapters I took a cognitivist or procedural view of language. The central assumption of this approach is the belief that by looking at the mechanisms which are used in language understanding and language production it will be possible to gain a clearer insight in the nature and structure of natural language. This thesis is an attempt to vindicate this belief. I developed a theory of the overall architecture of the human language processor and of the detailed structures which it uses in language production and constructs in language comprehension. I then used this theory in outlining analyses of various expressions in English. In this chapter I will discuss the major achievements and shortcomings of this thesis.

The thesis covers a broad range of subjects. Its main strength is the fact that in it a consistent theory of the human language processor is outlined based on the procedural view of language discussed in chapter 2. The usefulness of the model is shown in chapter 6 in which a number of analyses of linguistic expressions in English are sketched. The number of topics which are touched upon, is also responsible for the main weakness of the thesis:- although a large number of problems are discussed, very few of the solutions which are put forward, are worked out in any great detail. Especially the linguistic analyses in chapter 6 are very sketchy and should be done in more detail. An obvious way of getting around this problem would be to implement the proposals made in chapter 4 and 5. Although I have written a PROLOG program which "understands" descriptions of simple maps, no serious attempt has yet been made to do this.

In the remainder of this chapter I will discuss the various achievements and shortcomings of the various chapters. In it I will also point to possible directions of future research.

In chapter 2. I briefly outlined the procedural position in the semantics for natural language and showed that some of the ideas developed in naturalistic epistemology could be used to defend the cognitivist view of language against possible objections. In particular, I showed that the intersubjectivity of language and the fact that language can be used to talk about the world could both be accounted for if we accepted the thesis of cognitive similarity, the thesis that our cognitive systems are structurally more or less alike.

I argued that the cognitivist view has a number of consequences for theories of language. First, there were a number of consequences for criteria of adequacy which any satisfactory theory had to meet. Since language was to be studied by looking at the mechanisms underlying its use, I concluded that satisfactory theories had to meet both psycholinguistic criteria and linguistic criteria of adequacy. Secondly, the cognitivist view had implications for the form of a satisfactory (and complete) theory. Such a theory would make proposals about the overall architecture of the human language processor and about the more detailed structures which are constructed in language comprehension and used in language production.

The criteria of adequacy mentioned in chapter 2. were used in chapter 3 to discuss critically a number of relevant proposals in the literature. First, I argued against the Kintsch-Fodor programme of psychologizing formal semantics. I then reviewed the proposals for the architecture of the human language processor made by Sanford and Garrod, and by Johnson-Laird. Finally, I discussed a number of proposals about the structures people built up when understanding a piece of

discourse. One of the main conclusions of this chapter about the structure of the discourse representations in the human mind was that these representations were likely to be very similar in structure to the representations which were used to store knowledge in general. This of course implied that the work done in AI in the area of knowledge representation became crucial.

In chapter 4., I made a proposal for the overall architecture of the human language processor. The system I came up with was at least at first sight very similar to that of Sanford and Garrod, although the underlying view of the various components in the system was rather different. I proposed to see the human language processing mechanism as consisting of three different components which could be thought of as modelling different degrees to which knowledge could be activated. I distinguished between the discourse model, containing a representation of the utterance being understood or produced; the discourse specific epistemic model, containing the background knowledge relevant to the discourse, which came either from previous discourse or from the long-term knowledge store; and the general epistemic model, a component containing the background knowledge a processor would bring to the discourse. Another important idea introduced in this chapter was the idea of embedded models which were used to model the beliefs processors ascribed to their fellow interlocutors. I used this to sketch a theory of how to model mutual knowledge.

Since the model I propose is very much like the model of Sanford and Garrod, the same psychological evidence which they put forward can be put forward for mine. The model I propose avoids the difficulties which I raised in chapter 3. against Sanford and Garrod's model. It can thus be seen as an improvement on their model.

In chapter 5., I made a proposal for knowledge representation which was based on the idea of a frame as developed by Minsky and bore a close similarity to a knowledge representation scheme developed by Bobrow and Winograd. I argued for an organization of the knowledge base around conceptual entities with essential and default beliefs attached to them. I distinguished between two sorts of conceptual entities, mental objects and concepts. I provided a model-theory for the structures defined in this chapter, which made use of the notion of arbitrary objects developed by Kit Fine. This model-theory has the advantage of being more direct than alternative proposals in the literature. (e.g. Hayes, 1977).

I admitted that the formal knowledge representation system KRS was less expressive than other systems which are based on more or less the same intuitions. In particular, the only types of conceptual entity which were permitted, were mental objects and concepts. KRS has to be extended to allow for the representation of knowledge of relationships between objects, knowledge about (types of) events etc. Although it seems to me that syntactic extension of the system to allow for the representation of knowledge of this type is relatively straightforward, there are a number of problems from a semantic point of view. If one wants to maintain the direct semantic treatment for KRS, then Fine's theory of arbitrary objects will have to be extended to allow for arbitrary relationships and arbitrary events as well. How one would do this is an open question.

Chapter 6. finally was intended as an illustration of the viability of the cognitivist view of language. I first discussed the notion of reference from the procedural view. I then discussed various expressions in English, in particular indefinite NPs, definite descriptions, pronouns and 'universally' quantified NPs.

It is important to realize how the procedural view helped to clarify at least some of the problems associated with these expressions. First, I was able to explain the intuition that indefinites are used to introduce one and only one individual into the discourse without necessarily implying uniqueness in the world. Thus, we were able to square the existential quantifier analysis of indefinites with the intuition that indefinites introduce one and only one individual into the discourse. Secondly, we could show that the Russelian analysis of definite descriptions followed from the Strawsonian analysis given the processing principle of retrospective updating which was proposed on independent grounds. Obviously, without taking processing factors into account this would not have been possible. Thirdly, I explained the difference between pronouns and definite descriptions in terms of the degree to which the object which the hearer was supposed to access had been activated. The objects which were accessed by means of a pronoun existed in the discourse model, whereas those to be accessed by definites existed in the discourse specific epistemic model. The prediction which followed from this that all the examples found with pronouns had counterparts involving definite descriptions and the other way around, was borne out by the facts. Lastly, we were able to give an account of the universal quantifiers in processing terms. The non-standard quantifiers such as *most, few* could be explained in terms of the notion of default value, a notion which was introduced in chapter 5. on knowledge representation. It is important to point out that while the first and second point could be explained because we took the procedural view in general, the last two points followed from the proposals about the architecture of the human language processor and the proposals about the structure of the mental representations which are used or constructed in language production and understanding.

The linguistic analyses put forward are rather shallow. No attempt is made to analyse all occurrences of one particular type of expression in English. Given the fact that we are primarily interested in providing an overall theory of the human language processor, and the fact that the linguistic analyses are intended as illustrations of the usefulness of the particular proposals made here, this is maybe unavoidable. More detailed analyses of the expressions discussed here, and of other expressions in English, or any other natural language, would be useful since they might pinpoint weaknesses in the present system and suggest possible extensions and/or alterations.

Altmann, G. (forthcoming) *Reference and the resolution of local syntactic ambiguity: The effect of context during human sentence processing.* Unpublished PhD Thesis, University of Edinburgh.

Anderson, J. & G. Bower (1973) *Human associative memory.* Washington D.C.: Winston & Sons.

Barr, A. & E. Feigenbaum (eds.) *The handbook of artificial intelligence. Volume 1.* Los Altos, Cal: Kaufman.

Bartlett, F. (1932) *Remembering: A study in experimental and social psychology.* Cambridge: Cambridge University Press

Bobrow, D. & A. Collins (ed.) (1975) *Representation and understanding: Studies in cognitive science.* New York: Academic Press.

Bobrow, D. & M. Stefik (1983) *The LOOPS manual.* Intelligent Systems Laboratory. Xerox Corporation.

Bobrow, D. & T. Winograd (1977) An overview of KRL, a knowledge representation language. *Cognitive Science,* 1,3-46

Bosch, P. (1984) *Agreement and anaphora: A study of the role of pronouns in syntax and discourse.* New York: Academic Press.

Brachman, R. (1976) What's in a concept? Structured foundations for semantics networks. BBN Report 3433. Cambridge, Mass.: BBN.

Brachman, R. (1979) On the epistemological status of semantic networks. In Findler (1979).

Bransford, J. & M. Johnson (1972) Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behaviour.* 11, 717-26.

Bransford, J. & M. Johnson (1973) Consideration of some problems in comprehension. In W. Chase (ed.) *Visual information processing.* New York: Academic Press.

Brown, G & G. Yule (1984) *Discourse analysis.* Cambridge: Cambridge University Press.

Buehler, K. (1934) *Sprachtheorie.* Jena: Fischer.

Carnap, R. (1947) *Meaning and necessity.* Chicago: Chicago University Press.

Carlsson, G. (1977) A unified analysis of the English bare plural. *Linguistics and Philosophy,* 1, 413-457.

Chafe,W. (ed.) (1980b) *The pear stories: Cognitive, cultural and linguistic aspects of narrative production.* Norwood, N.J.: Ablex.

Chomsky, N. (1965) *Aspects of the theory of syntax.* Cambridge, Mass: MIT Press.

Christopherson, P. (1939) *The articles: A study of their theory and use in English.* Copenhagen: Munksgaard.

Clark, H. & T. Carlson (1982b) Speech acts and hearers' beliefs. In N. Smith (ed.) *Mutual knowledge.* New York: Academic Press.

Clark, H. & S. Haviland (1977) Comprehension and the given-new contract. In: R. Feedle (ed.) (1977)

Clark, H. & C. Marshall (1981) Definite reference and mutual knowledge. In A. Joshi et al (eds.).

Collins, A. & M. Quillian (1972) How to build a language user. In E. Tulving & W. Donaldson (eds.).

Crain, S. (1980) *Contextual constraints on sentence comprehension.* PhD thesis. University of Connecticut.

Crain, S. & M. Steedman (1983) On not being led up the garden path: The use of context by the psychological parser. In A. Zwicky & D. Dowty (eds.) *Natural language processing: Psychological, computational and theoretical perspectives* Cambridge: Cambridge Unversity Press

Crymes, R. (1968) *Some systems of substition correlations in modern American English.* The Hague: Mouton.

Dennett, D. (1978) *Brainstorms: Philosophical essays on mind and philosophy.* Hassocks: Harvester Press.

Donnellan, K. (1966) Reference and definite descriptions. *Philosophical Review,* 75, 281-304. (Also in Steinberg & Jakobovits (eds.) (1971).

Doyle, J. & Ph. London (1980) A selected descriptor-indexed bibliography to the literature on belief revision. MIT AI Memo 568. Also: SIGART Newsletter #71, April 1980.

Du Bois, J. (1980) Beyond definiteness: The trace of identity in discourse. In W. Chafe (1980) (ed.) *The pear stories: Cognitive, cultural and linguistic aspects of narrative production.* Norwood, N.J.: Ablex.

Ehlich, K. (1982) Anaphora and deixis: Same, similar or different. In: Jarvella and Klein (eds).

Fahlman, S. (1979) *NETL:- A system for representing and using real-world knowledge.* Cambridge Mass: The MIT Press.

Fauconnier, G. (1979) *Mental Spaces: A discourse processing view to natural language logic.* Unpublished paper, Universite de Paris.

Feedle, R. (ed.) (1977) *Discourse processes: advances in research and theory.* Norwood N.J.: Ablex

Fine, K. (1983) A defence of arbitrary objects. *Proceedings of the Aristotelian Society,* Supp vol LVII, 55-77.

Fine, K. (1985) Natural deduction and arbitrary objects. *Journal of Philosophical Logic,* 14, 57-107.

Findler, N. (ed.) (1979) *Associative networks: The representation and use of knowledge in computers.* New York: Academic Press.

Fodor, J.A. (1976) *The language of thought.* Hassocks: Harvester.

Fodor, J. & I. Sag (1982) Referential and quantificational indefinites. *Linguistics and Philosophy,* 5, 355-398.

Frege, G. (1892) Ueber Sinn und Bedeutung. *Zeitschrift fuer Philosophie under Philosophische Kritik,* 100,25-50.

Furukawa K., A. Takeuchi, S. Kunifuji, H. Yaukawa, M. Ohki & K. Ueda (1984) MANDALA: A logic based knowledge programming system. *Proceeedings of the international conference on fifth generation computer systems 1984.*

Garrod, S. & A. Sanford (1981) Bridging inferences and the extended domain of reference. In: J. Long & A. Baddeley (eds.) *Attention and performance.* Hillsdale, N. J.: Lawrence Erlbaum.

Garrod, S. & A. Sanford (1982) The mental representation of discourse in a focussed memory system: Implications for the interpretation of anaphoric noun phrases. *Journal of Semantics,* 1,21-41.

Goldberg, A. & D. Robson (1983) *Smalltalk-80, the language and its implementation.* Reading, Mass: Addison-Wesley.

Grosz, B.J. (1981) Focusing and description in natural language dialogue. In A. Joshi, B. Webber and I. Sag (eds.)

Haviland, S & H. Clark (1974) What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior,* 13, 512-21.

Hawkins, J. (1978) *Definiteness and indefiniteness: A study in reference and grammaticality predication.* London: Croom Helm.

Hawkins, J. (1982) The definite article and a structured universe of discourse. Unpublished manuscript: Max Planck Gesellschaft, Nijmegen.

Hayes, Patrick (1977) The logic of frames. In D. Metzing (ed.) *Frame conceptions and text understanding.* Berlin; de Gruyter.

Heim, I. (1982) *The semantics of definite and indefinite noun phrases.* Doctoral dissertion submitted at the University of Massachusetts.

Hendrix, G. (1975) Expanding the utility of semantic networks through partitioning. Stanford Research Institute.

Hendrix, G. (1979) Encoding knowledge in partioned networks. In Findler (1979)

Hewitt, C. (1977) Viewing control structures as patterns of message passing. *Artificial Intelligence,* 8, 324-64.

Israel, D. & R. Brachman (1984) Some remarks on the semantics of representation languages. In M. Brodie, J. Mylopoulos & J. Schmidt (eds.) *On conceptual modelling: Perspectives from artificial intelligence, databases and programming languages.* Berlin: Springer Verlag.

Jarvella, R. & W. Klein (ed.) (1982) *Speech, place and action.* Chichester: Wiley.

Johnson-Laird, Ph. (1982) Formal semantics and the psychology of meaning. In S. Peters & E. Saarinen (eds.) *Processes, beliefs and questions.* Dordrecht: Reidel.

Johnson-Laird, P. (1983) *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge: Cambridge University Press.

Johnson-Laird, Ph. & A. Garnham (1980) Descriptions and discourse models. *Linguistics and Philosophy,* 3,371-93.

Johnson-Laird, P & M. Steedman (1978) The psychology of syllogisms. *Cognitive Psychology,* 10,64-99.

Joshi, A., B. Webber & I. Sag (eds.) (1981) *Elements of discourse understanding.* Cambridge: Cambridge University Press.

Kamp, H. (1981) A theory of truth and semantic representation. In: J. Groenendijk, T. Janssen & M. Stokhof (eds.) *Formal methods in the study of language.* Amsterdam: Mathematisch Centrum.

Karttunen, L. (1969) Pronouns and variables. *CLS,* 5.

Karttunen, L. (1976) Discourse Referents. In: J. McCawley (ed.) *Syntax and semantics, volume 7.* New York:Academic Press.

Kintsch, W. (1974) *The representation of meaning in memory.* Hillsdale, N.J.: Lawrence Erlbaum.

Kuijpers, B. (1975) A frame for frames: Representing knowledge for recognition. In Bobrow & Collins (1975).

Kuno, S. (1975) Three perspectives in the functional approach to syntax. In R. Grossman, L. San & T. Vance (eds) *Functionalism.* Papers from the parasession on functionalism. Chicago Linguistic Society.

Kuroda, S-Y. (1969) English relativization and certain related problems. In D. Reibel & S. Schane (eds) *Modern Studies in English.* Englewood Cliffs, N.J.: Prentice Hall.

Lakoff, G. (1982) Experimental factors in linguistics. In Th. Simon & R. Scholes (eds.) (1982).

Lewis, D. (1969) *Convention: A philosophical study.* Cambridge, Mass.: Harvard University Press.

Locke, J. (1690) *An inquiry concerning human understanding.*

Lyons, J. (1979) Deixis and anaphora. In T. Myers (ed.) *The development of conversation and discourse.* Edinburgh: Edinburgh University Press.

Mani, K. & P. Johnson-Laird (1982) The mental representation of spatial descriptions. *Memory and Cognition,* 10,181-7

Marslen-Wilson, W., E. Levy & L. Tyler (1982) Producing interpretable discourse: The establishment and maintenance of reference. In Jarvella and Klein (1982).

McAllister, D. (1980) An outlook on truth maintenance. MIT AI Memo No. 551, August 1980.

McCawley, J. (1979) Presupposition and discourse structure. In: C. Oh & D. Dinneen (eds.) *Syntax and semantics 11: Presupposition.* New York: Academic Press.

McCawley, J. (1982) How far can you trust a linguist. In Th. Simon & R. Scholes (eds.) (1982)

Minsky, M. (1975) A frame work for representing knowledge. In P. Winston (ed.) *The psychology of computer vision.* New York: McGraw Hill.

Montague, R. (1970a) English as a formal language. In B. Visentini et al. *Linguaggi nella societa e nella tecnica.* Milan: Edizioni di Communita. (also in Thomason (ed.) (1974))

Montague, R. (1970b) Universal grammar. *Theoria,* 36,373-90 (also in Thomason (ed.) (1974))

Montague, R. (1973) The proper treatment of quantification in English. In: J. Hintikka, J. Moravscik & P. Suppes (eds.) *Approaches to natural language: Proceedings of the 1970 Stanford workshop in grammar and semantics.* Dordrecht: Reidel (also in Thomason (ed.) (1974))

Prince, E. (1981) On the inferencing of indefinite- *this* NPs. In Joshi, Webber and Sag (eds.)

Reichgelt, H. (1981) *Ego in semantics: Two studies in speaker orientation.* Unpublished M.A. Thesis, University of Nijmegen.

Reichgelt, H. (1982) Mental models and discourse. *Journal of Semantics,* 1,371-86.

Richards, B. (1984) On interpreting pronouns. *Linguistics and Philosophy,* 7, 287-324.

Russell, B. (1905) On denoting. *Mind,* 14,479-93.

Said, K. (1985) Axiomatic epistemic logics and artificial intelligence. *Proceedings of AISB 85.* Coventry: University of Warwick.

Sanford, A & S. Garrod (1981) *Understanding written language: Explorations beyond the sentence.* Chichester: John Wiley & Sons.

Sanford, A., S. Garrod, A. Lucas & R. Henderson (1983) Pronouns without explicit antecedents? *Journal of Semantics,* 2, 303-18.

Schiffer, S. (1972) *Meaning.* Oxford: Clarendon.

Searle, J. (1969) *Speech acts: An essay in the philosphy of language.* Cambridge: Cambridge University Press.

Seuren, P. (1975) *Tussen taal an denken: Een bijdrage tot de empirische funderingen van de semantiek.* Utrecht: Oosthoek, Scheltema & Holkema.

Seuren, P. (1982) *The construction of discourse domains through accumulated increments.* Paper delivered at Edinburgh conference on language, reasoning and inference.

Seuren, P. (1985) *Discourse semantics.* Oxford: Blackwell.

Shadbolt, N. (1983) Processing reference. *Journal of Semantics*, 2,63-98.

Sidner, C. (1983) Focusing for the interpretation of pronouns. *American Journal of Computational Linguistics*, 7, 217-31.

Simon, Th. & R. Scholes (eds.) (1982) *Language, mind, and brain.* Hillsdale, New Jersey:Lawrence Erlbaum.

Shadbolt, N. & H. Reichgelt (forthcoming) *Reference in discourse.*

Soames, S. (1984) Linguistics and psychology. *Linguistics and Philosophy*, 7, 155-80.

Stalnaker, R. (1970) Pragmatics. *Synthese*, 272-289 (Also in Davidson & Harman (eds.) (1972))

Stalnaker, R. (1978) Assertion. In P. Cole (Ed.) (1978) *Syntax and semantics 9: Pragmatics.* New York: Academic Press.

Steinberg, D. & L. Jakobovits (eds.) (1971) *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology.* Cambridge: Cambridge University Press.

Stenning, K. (1977) Articles, quantifiers and their encoding in textual comprehension. In: R. Feedle (ed.) (1977)

Stenning, K. (1981) On remembering how to get there: How one might want something like a map. In A. Lesgold, J. Pelligrino, S. Fokkema, & J. Glaser (eds.) *Cognitive psychology and instruction.* New York: Plenum.

Stirling, L. (1981) *Anaphora and deixis in doctor-patient consultations.* Unpublished dissertation. University of Queensland

Strawson, P. (1950) On referring. *Mind*, 59,320-44. (Also in Strawson (1971)).

Strawson, P (1964) Identifying reference and truth values. *Theoria*, 30 (Also in Strawson, 1971).

Strawson, P. (1971) *Logico-linguistic papers.* London: Methuen.

Sulin, R. & D. Dooling (1974) Intrusion of a thematic idea in retention of prose. *Journal of Experimental Psychology,* 103,255-62

Thomason, R. (ed) (1974) *Formal philosophy: Selected papers of Richard Montague.* New Haven: Yale University Press.

Troelstra, A. (1969) *Principles of intuitionism.* Berlin: Springer-Verlag.

Tulving, E. & W. Donaldson (eds.) (1972) *Organization of memory.* New York: Academic Press.

Van Benthem, J. and J. van Eijck (1982) The dynamics of interpretation. *Journal of Semantics,* 1,3-20.

Webber, B. (1978) *A formal approach to discourse anaphora.* BBN report 3761. Cambridge, Mass.: BBN.

Winograd, T. (1976) Towards a procedural understanding of semantics. *Revue International de Philosophie,* 3,260-303.

Yule, G. (1982), Interpreting anaphora without identifying reference. *Journal of Semantics,* 1, 315-22.